eLife

Neuroscience
Computational and Systems Biology

Gain neuromodulation mediates task-relevant perceptual switches: evidence from pupillometry, fMRI, and RNN Modelling

Reviewed Preprint
v2 • February 27, 2025
Revised by authors

Reviewed Preprint
v1 • January 25, 2024

Gabriel Wainstein, Christopher J Whyte, Kaylena A Ehgoetz Martens, Eli J Müller, Vicente Medel, Britt Anderson, Elisabeth Stöttinger, James Danckert, Brandon R Munn, James M Shine

Brain and Mind Center, The University of Sydney, Sydney, Australia • Center for Complex Systems, The University of Sydney, Sydney, Australia • The University of Waterloo, Waterloo, Canada • Latin American Brain Health (BrainLat), Universidad Adolfo Ibanez, Santiago, Chile • Hochschule Fresenius, Idstein, Germany

eLife Assessment

This **valuable** paper explores the idea that transient modulations of neural gain promote switches between distinct perceptual interpretations of ambiguous stimuli. The authors provide **solid** evidence for this idea by pupillometry (an indirect proxy of neuromodulatory activity), fMRI, neural network modeling, and dynamical systems analyses. The highly integrative nature of this approach is rare in the field.

https://doi.org/10.7554/eLife.93191.2.sa3

Abstract

Perceptual updating has been hypothesized to rely on a network reset modulated by bursts of ascending neuromodulatory neurotransmitters, such as noradrenaline, abruptly altering the brain's susceptibility to changing sensory activity. To test this hypothesis at a large-scale, we analysed an ambiguous figures task using pupillometry and functional magnetic resonance imaging (fMRI). Behaviourally, qualitative shifts in the perceptual interpretation of an ambiguous image were associated with peaks in pupil diameter, an indirect readout of phasic bursts in neuromodulatory tone. We further hypothesized that stimulus ambiguity drives neuromodulatory tone leading to heightened neural gain, hastening perceptual switches. To explore this hypothesis computationally, we trained a recurrent neural network (RNN) on an analogous perceptual categorisation task, allowing gain to change dynamically with classification uncertainty. As predicted, higher gain accelerated perceptual switching by transiently destabilizing the network's dynamical regime in periods of maximal uncertainty. We leveraged a low-dimensional readout of the RNN dynamics, to develop two novel macroscale predictions: perceptual switches should occur with peaks in low-dimensional brain state velocity and with a flattened egocentric energy landscape. Using fMRI we confirmed these predictions, highlighting the role of the neuromodulatory system in the large-scale network reconfigurations mediating adaptive perceptual updates.

Introduction

The overwhelming majority of neurons in our brains have only indirect interactions with the external world. This means that the identity of sensory inputs is inherently ambiguous[1–5]. The equivocal nature of perceptual input is overcome by incorporating prior information about the causal structure of the world into sensory inferences. This is clearly evidenced in laboratory experiments that present participants with sensory inputs that offer two equally valid yet mutually exclusive perceptual interpretations (e.g. the Necker cube illusion and binocular rivalry): in these ambiguous scenarios, observers periodically switch between mutually exclusive percepts[6–8].

Outside of conditions of extreme perceptual ambiguity, perceptual awareness is remarkably stable, suggesting that the nervous system can rapidly (and flexibly) identify the best 'match' between visual data and a stable (likely known) stimulus category[6,9]. Importantly, this process of combining ambiguous sensory input with prior information must be dynamic: adaptive behaviour requires that the relative reliability of prior information and current sensory input are made suitably contextually dependent[10–13]. In ecological settings, the problem is even more pronounced: not only does the reliability of the sensory input vary, the urgency of perceptual decision making also changes between context[14–16].

Neuroimaging studies investigating perceptual updating and switches have typically identified a distributed set of regions within the cerebral cortex[17,18]. These cortical regions are presumed to play a role in attentional shifts driving switches in perceptual contents by selectively boosting activity within the relevant circuits[18,19]. This interpretation is complemented by behavioural evidence showing that attention plays a prominent role in determining the contents of perception in bistable perception tasks where competition is not resolved at low-levels of the visual hierarchy[21,22]. Similarly, computational models of perceptual decision making typically consist of winner-take-all competition between cortical populations[23–25]. Yet, the ability to flexibly respond to ambiguous visual inputs according to changing task demands is a feature that is present across phylogeny[26] and hence is present in a wide variety of animals that have poorly developed cerebral cortices[27]. Indeed, phasic change in the highly conserved ascending arousal system have been linked to moment-by-moment adaptive updates in the relative weighting of prior information, sensory input, and the urgency of the perceptual decision process through neuromodulatory mediated alterations in neural gain[28–31].

The ascending neuromodulatory system, and specifically the noradrenergic locus coeruleus (LC), is well-suited to modulate the large-scale, brain state switches required to flexibly alter perceptual contents[32,33]. While the cell body of the LC is located in the brainstem, the nucleus sends projections throughout the central nervous system, wherein its axons release noradrenaline, which in turn modulate the excitability of targeted regions[11]. In previous work, it has been argued that the phasic release of noradrenaline from the LC acts as a "network reset" signal, which effectively disrupts ongoing processing, and hence allows animals to reconfigure their ongoing neural dynamics towards more salient (and hopefully, behaviourally-relevant) processes[34–36]. This mechanism is of critical importance in ecological contexts in which an animal needs to be able to both focus on the current task in an exploitative mode (such as foraging), while being able to rapidly modify its internal, attentional and behavioural state when required (e.g., if resources are depleted, or in the presence of a predator).

Preliminary evidence in the context of bistable perception has shown that when a stimulus is task-relevant, pupil diameter (a non-specific and indirect readout of phasic LC activity[37–40] and neuromodulatory tone) is tightly linked to switches in the content of perception[31,41,42]. In line with this, recent modelling has shown that linking perceptual updates to fluctuations in

neuromodulatory tone recapitulates the phasic-tonic firing rate pattern known to characterise LC spiking dynamics and improves performance in reinforcement learning tasks[28]. Thus, whilst the LC could plausibly mediate perceptual switches in a task-relevant setting, we still need a more robust test of this hypothesis.

Based on previous work[34–36,43,44], and the projections of the LC to many of the regions implicated in whole-brain imaging studies of perceptual uncertainty[38,45,46], we hypothesized that task-related perceptual switches can be modulated by phasic bursts of LC activity, which act as a 'network reset'[36], flattening the whole-brain energy landscape[47] and thus allowing cortical dynamics to evolve into a new state thereby changing the contents of perception.

To test this hypothesis, we leveraged a cognitive task designed to investigate switches in perceptual categorisation[48]. We observed that pupil diameter peaked at the point of the perceptual switch and predicted their timing. We then trained a recurrent neural network (RNN) to perform an analogous change detection task. Based on previous modelling and theory we allowed the gain of the activation function (an established mechanism for the action of noradrenaline on the cerebral cortex[11,49,50]) to vary as a function of the uncertainty in the pretrained network's perceptual categorisation. This revealed that heightened gain facilitated earlier perceptual switches by transiently destabilizing the network's dynamics under conditions of maximal uncertainty. Further analyses translated these neural dynamics into two predictions that could be tested in fMRI data[17,48]: (1) heightened gain increases the velocity of low-dimensional neural trajectories around perceptual switches, and (2) it flattens the energy landscape of the neural state space[47]. Overall, our results support the hypothesis that phasic bursts of neuromodulatory activity act as a "network reset" [34,36], dynamically disrupting stable network states and facilitating switches in perceptual categorisation. This reset mechanism highlights the role of neuromodulatory systems in transiently reorganizing network dynamics to enhance flexibility and adaptability in response to uncertainty.

Results

Evoked pupil dilations coincide with the resolution of perceptual ambiguity

To assess the role of the ascending arousal activity during task performance, we analysed a dataset of 35 participants who performed an ambiguous figures task whilst simultaneously recording pupil diameter with an eye tracker device (SR Research, 1000 Hz). Briefly, the task consisted of a set of continuously transforming images that transition from an initial object (e.g., a shark) into a second object (e.g., a plane), while preserving basic psychophysical attributes (**Fig. 1A**). Crucially, even though the task stimuli change incrementally and linearly, with maximal ambiguity at the mid-point of each trial (the peak of the dotted line curve in **Fig. 1A**), awareness of a change in the stimulus is known to 'pop out', often at different times on each trial[48]. When these perceptual switches occurred, subjects were instructed to change the button they were pressing, thus indicating a change in perceptual interpretation across stimuli. Participants viewed 20 unique sets of images, each of which morphed from a starting image into a second image through 15 equally spaced intermittent stages (**Fig. 1A**). For each participant, we identified the first and last time they viewed a sequence of images, as well as the three images leading up to and following an identified perceptual switch, irrespective of the categories associated with each specific object switch. The rest of the analyses in this manuscript are organised around this perceptual transition.

Given the known (admittedly non-specific) relationship between LC activity and the dynamics of pupil diameter[51] (**Fig. 1B**), we were able to test the hypothesis that neuromodulatory tone is associated with perceptual switching. The linear nature of the morphing procedure meant that luminance levels (which could otherwise bias pupil diameter[32,37,38]) were kept constant across all trials. Additionally, motor preparation was controlled by requiring subjects to press a button on each image (indicating the content of their perception). Mapping all blink-corrected, filtered and normalized trials over time.

We observed a clear increase in the phasic pupillary response approximately three trials before participants switched to a new perceptual category, potentially reflecting the onset of increased ambiguity toward a new object (**Fig. 1C**). This response peaked at the point of the perceptual switch, corresponding to the maximum pupil diameter (**Fig. 1C**). Further analysis revealed a significant increase in the mean pupil response starting three images before the change point (mean $\beta = 0.22$; $t_{(32)} = 8.02$, $p = 2.3 \times 10^{-19}$), before returning to baseline levels.

Next, we sought to elucidate the relationship between ascending arousal, quantified by pupil diameter, and the temporal dynamics of perceptual shifts on a trial-by-trial basis. Given the pivotal role of the LC in modulating sensory processing and perceptual switches (**Fig. 1B**), we hypothesized that the speed of a perceptual switch would correlate with neuromodulatory tone. Specifically, we predicted that trials with faster perceptual switches would be associated with an increase in pupil diameter, while slower switches would correspond to a decrease.

To test this prediction, we performed a two-level linear model analysis. The peak pupil diameter observed during the perceptual switch was designated as the independent variable, and the trial on which the perceptual shift was reported served as the regressor for each subject. To control for potential confounds, such as impulsive premature responses, and to address reduced statistical power in extreme response epochs (both early and late), we limited our analysis to responses within two images from the median switch point (9 ± 2; 84.1% of total trials). At the group level, we conducted a one-tailed t-test on the regressors from the linear model. As expected, we observed an inverse relationship between evoked pupil diameter and the trial marking the perceptual switch (mean $\beta = -0.19$, $t_{(27)} = -2.6452$, $p = 6.7 \times 10^{-3}$, SD = 0.3880). Earlier responses showed a positive relationship with higher evoked pupil diameter during the switch epoch (**Fig. 1D** red), whereas later responses were associated with a more constricted pupil (**Fig. 1D** blue). In summary, these results provide indirect evidence for our hypothesis that ascending neuromodulation – such as LC activity – is associated with the speed of perceptual switches.

Computational evidence for neuromodulatory-mediated perceptual switches in a recurrent neural network

Our initial results provided confirmatory evidence implicating the neuromodulatory tone of the ascending arousal system in perceptual switches. There is evidence, however, suggesting that simply changing stimulus categories can also induce similar pupillary dilations[41,52,53]. What we need, therefore, is a more mechanistic means of both framing and testing our network reset hypothesis in the context of perceptual switching. Along with others[54–57], we have used a combination of computational modelling[49,58], neurobiological theory[11], and multi-model neuroimaging[19,27,31,32] to suggest that noradrenaline alters neural gain[50,61], which in turn affects inter-regional communication flattening the energy landscape traversed by the brain's dynamics allowing the brain state to jump between perceptual attractors more easily. Whether these signatures of large-scale network reconfiguration are mechanistically related to network reset remains an important and open question.

To test whether our hypothesised neuromodulatory mechanism could recapitulate the behaviour we observed in the ambiguous figures task, we trained 50 continuous time recurrent neural networks (RNN) constrained to respect Dale's law (i.e., 80/20 split of purely excitatory/inhibitory

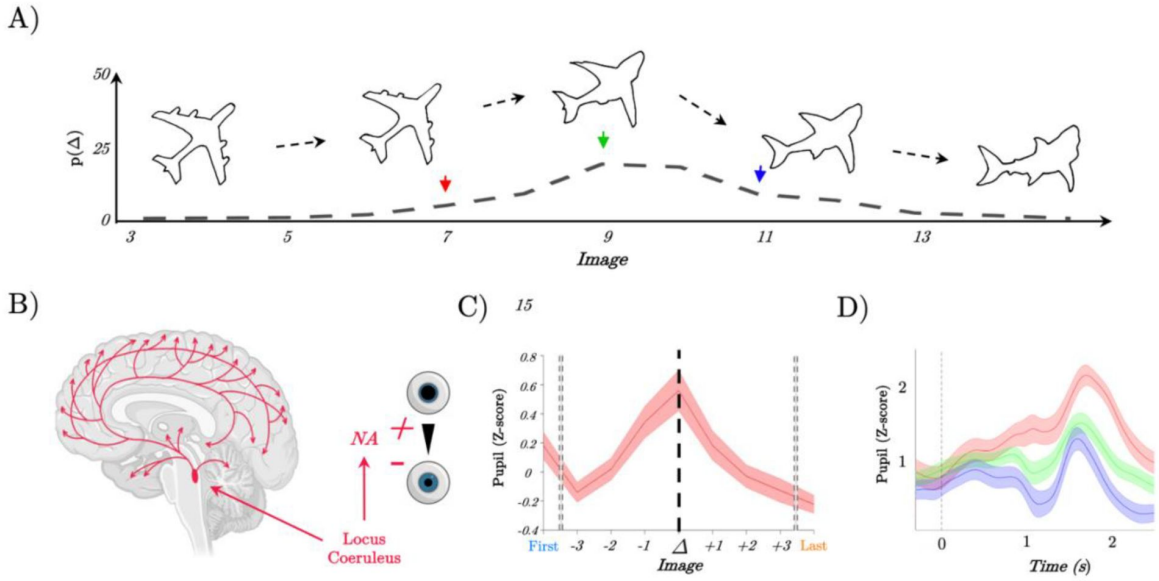


Figure 1

Pupil diameter tracks perceptual change.

A) Example trial showing the continuous change from a stable image (plane) into a shark; Lower: the probability of detecting a switch (*Δ*) as a function of Image – most switches occur around the mid-point, but not exclusively so, leading to our prediction of heightened locus coeruleus activity at the switch point; B) Representation of the locus coeruleus (red), its diffuse projections to the whole brain network and its link to pupil dilation. C) Pupil diameter group average evoked response time locked to the perceptual change (dark line, t = 0), significance is shown in the top grey bar ($p_{FDR} < 0.05$), showing two images around the perceptual change (*Δ*) are different from the null model. The average of the first and last two images are shown in the left (right) section of the plot (dotted line). We observed an increase of the pupillary response that peaked after the perceptual change. D) Group average of evoked pupillary responses to image switches – red represents the faster response when the switch occurs at image 6; green indicates a medium response with the switch at image 8; and blue denotes the slowest response with the switch at image 10.

units[62,63]) to perform a perceptual change detection task analogous to the task performed by our participants (**Fig. 2A**). The input and readout weights were constrained to be purely excitatory and only the firing rate of excitatory units contributed to the readout[64] (see Materials and Methods).

Each network was provided with a two-dimensional input $u(t) = [u_1\ u_2]^T$, with each column representing the "sensory evidence" for each of the two stimulus categories (**Fig. 2A**). The task lasted for 1 second of simulation time (we used a shorter time period for the simulation than the empirical task so that we could keep the integration step relatively small making the training and simulations more numerically tractable): to mimic the linear transition between image categories in our task, each trial began with maximum evidence for one of the two categories and minimum evidence for the other (e.g., $u_1 = 1$,$u_2 = 0$), and then linearly changed the evidence over the course of each trial such that by the final time-step the evidence for each category had switched (e.g., $u_1 = 0$,$u_2 = 1$). At each time point, the network was trained to output a categorical response indicating which input dimension had a higher value (**Fig. 2B**). Following training, all networks achieved near perfect behavioural accuracy (0.97 ± 0.02).

We next sought to test our hypothesis about the role of neural gain in perceptual switches. In previous work, we (and others) have argued that the impact of neuromodulators (such as NA) on population-level activity can be approximated by steepening (or flattening) the sigmoid activation function, thus mimicking the effect NA has on neuronal excitability by liberating intracellular calcium stores and/or opening (or closing) voltage-gated ions channels[11,65]. As a first test of this hypothesis, we manipulated the *gain* of the sigmoid activation function for all units in the network across a range of gain values (0.5 to 1.5) in a static manner. As predicted, increased gain (red; corresponding to heightened adrenergic tone) lead to earlier 'perceptual switches' in the network output whereas low gain caused later switches (**Fig. S1**).

Having confirmed that static manipulations of gain alter the speed of perceptual switches we constructed a more precise test of our hypothesis. Specifically, inspired by previous theoretical and experimental work showing that sensory prediction errors (i.e. transient increases in perceptual uncertainty) lead to phasic bursts in the noradrenergic locus coeruleus[28,30] we made gain time dependent with dynamics governed by a linear ODE with a forcing term proportional to the uncertainty (i.e. the entropy $H(z) = \sum_i p(z)_i \ln(p(z)_i))$ of the network's readout (**Fig. 2A**).

$$\tau \frac{dg}{dt} = (g_{tonic} - g) + \gamma H(z)$$

When the network's readout becomes uncertain approaching the perceptual switch (i.e. has high entropy) gain increases in a phasic manner (with magnitude γ), and in absence of the forcing, gain decays exponentially to its tonic value ($g_{tonic} = 1$). This modification resulted in gain dynamics reminiscent of the participant's pupil-diameter (**Fig. 2D**), and crucially, the speed of perceptual switches increased with the magnitude of the uncertainty driven forcing term (γ; **Fig. 2F**).

Having confirmed our hypothesis that increasing gain as a function of the network uncertainty increased the speed of perceptual switches, we next sought to understand the mechanisms governing this effect starting with the circuit level and working our way up to the population level (c.f. Sheringtonian and Hopfieldian modes of analysis[66]). Because of the constraint that the input and output weights were strictly positive, we could use their (normalised) value as a measure of stimulus selectivity. Inspection of the firing rates sorted by input weights revealed that the networks had learned to complete the task by segregating both excitatory and inhibitory units into two stimulus-selective clusters (**Fig. 2C**). As the inhibitory units could not contribute to the

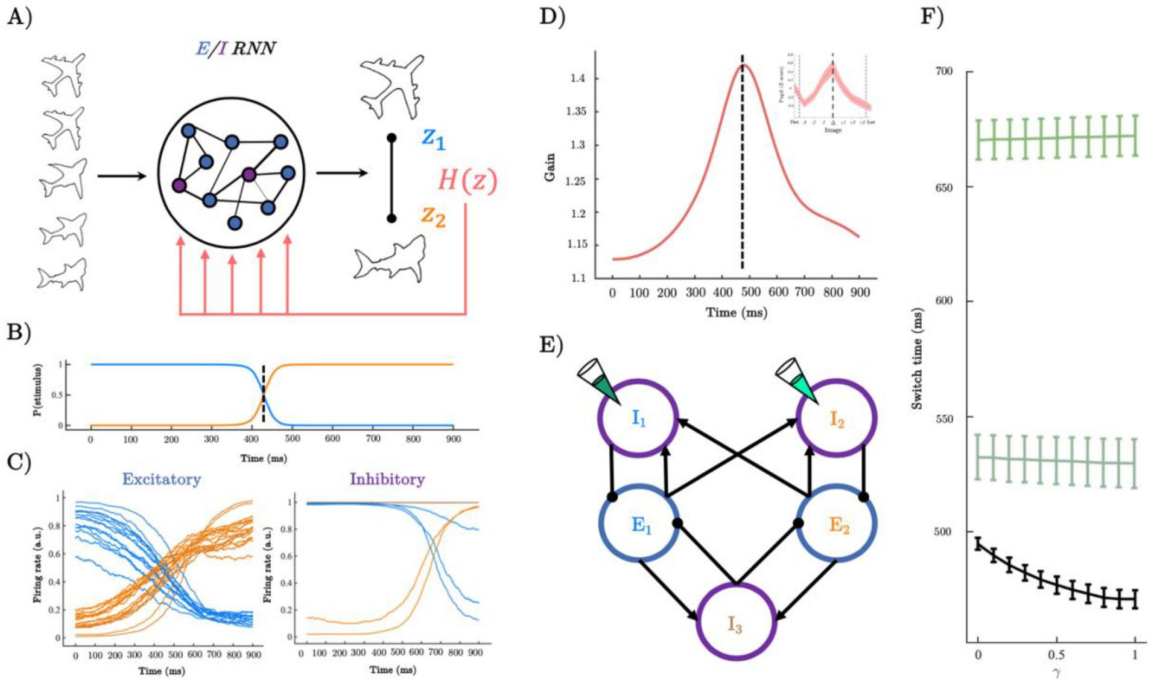


Figure 2

A recurrent neural network model of perceptual switching.

A) we trained a continuous time E/I recurrent neural network (RNN) to categorise linearly changing inputs representing two discrete categories (e.g., output z_1 and output z_2). B) Softmax of network outputs on example trial with γ = *.6*, dotted line shows the timing of the perceptual switch. C) Following training, the firing rate of the excitatory units was clearly separated into two stimulus selective clusters - those that responded maximally to u_1 (blue) and those that respond maximally to u_2 (orange). Inhibitory units demonstrated a similar modular clustering but were sorted by the selectivity of the excitatory units they inhibited. D) Dynamics of gain on example trial with γ = *.6* which peaks close to the perceptual switch (inset shows similarity to pupil diameter). E) Simplified network structure implied by selectivity analysis. Excitatory units (blue) form two stimulus selective modules. Each excitatory cluster is inhibited by a cluster of inhibitory units and a third non-selective inhibitory population. Pipette show lesion targets. F) Switch time as a function of γ magnitude (i.e. magnitude of uncertainty forcing). Lower black line shows a speeding effect of heightened γ (and therefore heightened gain at the perceptual switch). Teal lines show switch time for lesions to the inhibitory population targeting the initially dominant population (dark teal upper), and lesions to the inhibitory the population selective for the stimulus the input is morphing into (light teal middle).

networks read out, we hypothesised that they likely played an indirect role in perceptual switching by inhibiting the population of excitatory neurons selective for the currently dominant stimulus allowing the competing population to take over and a perceptual switch to occur.

To test this hypothesis, we sorted the inhibitory units by the selectivity of the excitatory units they inhibit (i.e. by the normalised value of the readout weights). Inspecting the histogram of this selectivity metric revealed a bimodal distribution with peaks at each extreme strongly inhibiting a stimulus selective excitatory population at the exclusion of the other (**Fig. S2**). Based on the fact that leading up to the perceptual switch point both the input and firing rate of the dominant population are higher than the competing population, we hypothesized that gain likely speeds perceptual switches by actively inhibiting the currently dominant population rather than exciting/disinhibiting the competing population. We predicted, therefore, that lesioning the inhibitory units selective for the stimulus (i.e. with normalised selectivity > *0.5*) that is initially dominant would dramatically slow perceptual switches, whilst lesioning the inhibitory units selective for the stimulus the input is morphing into would have a comparatively minor slowing effect on switch times since the population is not receiving sufficient input to take over until approximately half way through the trial irrespective of the inhibition it receives. As selectivity is not entirely one-to-one, we expect both lesions to slow perceptual switches but differ in magnitude. In line with our prediction, lesioning the inhibitory units strongly selective for the initially dominant population greatly slowed perceptual switches (**Fig. 3F** upper), whereas lesioning the population selective for the stimulus the input morphs into removed the speeding effect of gain but had a comparatively small slowing effect on perceptual switches (**Fig. 3F** lower).

Having found a circuit level explanation for the speeding effect of gain we next sought to understand the network's behaviour at a population level by interrogating the parameter space (with dimensions defined by network input and gain) traversed by the network. Unlike standard non-linear dynamical systems with stationary or (very) slowly time varying parameters, input and gain change rapidly over the course of each trial dynamically shifting the location and existence of the attractors shaping the network dynamics. Each trial is, therefore, characterised by a trajectory through a two-dimensional parameter space with dimensions corresponding to the gain of the activation function and the mismatch between input dimensions (Δinput).

Based on the selectivity of the network firing rates, we hypothesised that the dynamics were shaped by a fixed-point attractor, whose location and existence were determined by gain and Δinput, and changed dynamically over the course of a single trial[67–70]. Because of the large size of the network, we could not solve for the fixed points or study their stability analytically. Instead, we opted for a numerical approach and characterised the dynamical regime (i.e. the location and existence of approximate fixed-point attractors) across all combinations of gain and Δinput visited by the network. Specifically, for each combination of elements in the parameter space $\theta \in \mathbb{R}^{\text{gain} \times \Delta\text{input}}$ we ran 100 simulations with initial conditions (firing rates) drawn from a uniform distribution between [0,1], and let the dynamics run for 10 seconds of simulation time (10 times the length of the task - longer simulation times did not qualitatively change the results) without noise. As we were interested in the existence of fixed-point attractors rather than their precise location, at each time point we computed the difference in firing rate between successive time points $(\Delta r = \sum_i r_i(t) - r_i(t - \Delta t))$ across the network. For each simulation we computed both the proportion of trials that converged to a value of Δr below 10^{-2} giving us proxy for the presence of fixed points, and the time to convergence, giving us a measure of the "strength" of the attractor.

Across gain values when Δinput had unambiguous values ($u_1 \gg u_2$ or $u_2 \gg u_1$), the network rapidly converged across all initialisations (**Fig. 3A & 3C-H**). When Δinput became ambiguous, however, the dynamics acquired a decaying (inhibition-driven) oscillation and on many trails did

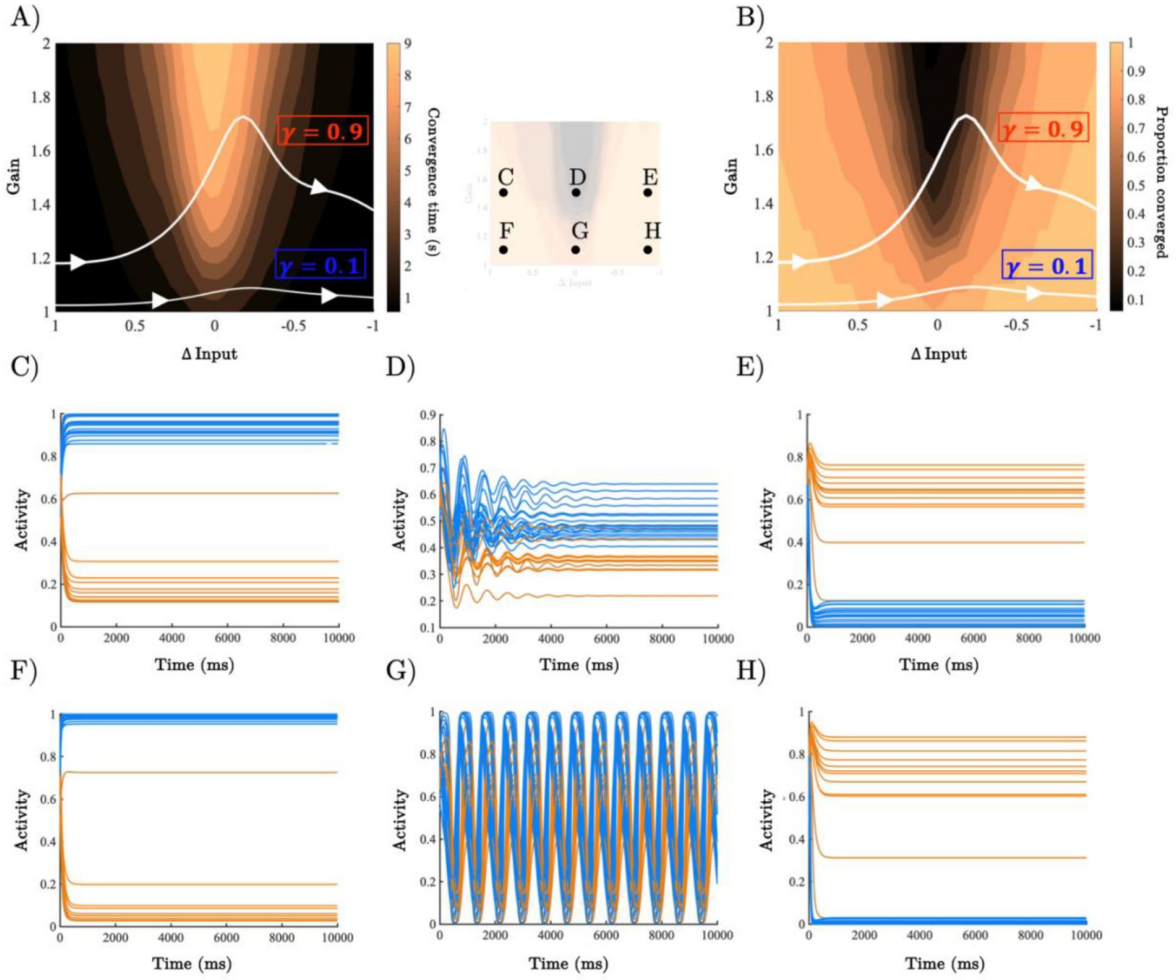


Figure 3.

Analysis of RNN dynamical regime.

A) Contour map of convergence time across the full gain by Δinput parameter space averaged across 100 initialisations with random initial conditions. Example parameter trajectories shown in white for high and low γ trials. B) Contour map of convergence proportion across the full parameter space. C-E) Example dynamics with gain = 1.1 and Δinput ≈ [1,0], [.5, .5], and [*0, 1*] respectively. F-H) Example dynamics with gain = 1.5 and Δinput ≈ [1,0], [.5, .5], and[*0, 1*] respectively.

not converge within the time frame of the simulation. As gain increased, the range of Δinput values characterised by oscillatory dynamics broadened. Crucially, for sufficiently high values of gain, ambiguous Δinput values transitioned the network into a regime characterised by high amplitude oscillations (**Fig. 3D & 3G**). Each trial can, therefore, be characterised by a trajectory through this 2-dimensional parameter space, with dynamics shaped by the dynamical regimes of each location visited (**Fig. 3A-B**).

When uncertainty had a small impact on gain (low γ) the network had a trajectory through an initial regime characterised by the rapid convergence to a fixed point where the population representing the initial stimulus dominated whilst the other was silent (**Fig. 3C**), an uncertain regime characterised by oscillations with all neurons partially activated (**Fig. 3D**), and after passing through the oscillatory regime, the network once again entered a (new) fix-point regime where the population representing the initial stimulus was silent whilst the other was dominant (**Fig. 3E**).

For high γ trails, the network again started and finished in states characterised by rapid convergence to a fixed point representing the dominant input dimension (**Fig. 3F-H**). However, it differed in how it transitioned between these states. Uncertain inputs generated high amplitude oscillations, causing the network to flip-flop between active and silent states (**Fig. 3G**). We hypothesised that, within the task, this mechanism silenced the initially dominant population, while boosting the competing population. To test this, we initialised each network with parameter values well inside the oscillatory regime (u ≈ [*.5 .5*], gain = 1.5) with initial conditions determined by the selectivity of each unit. Excitatory units selective for u_1, as well as the associated inhibitory units projecting to this population, were fully activated, whilst the excitatory units selective for u_2 (and the associated inhibitory units) were silenced (and vice versa for $u_2 \rightarrow u_1$ trials). As we predicted, when initialised in this state the network dynamics displayed an out of phase oscillation where the initially dominant population was rapidly silenced and the competing population was boosted after a brief delay (219 (ms), ± *114*; **Fig. S3**).

At the population level, therefore, heightened gain at points of ambiguity accelerates perceptual switches by transiently pushing the dynamics into an unstable regime. This regime replaces the fixed-point attractor representing the input with an oscillatory regime that actively inhibits the currently dominant population and boosts the competing population, before transitioning back to a stable (approximate) fixed-point attractor representing the new stimulus (**Fig. 3F-H** & **Fig. S3**).

Large-scale neural predictions of recurrent neural network model

Having confirmed the behavioural component of our gain modulation hypothesis in our model, and characterised both the circuit and population level mechanisms, we next sought to test our hypotheses that the speeding effect of uncertainty driven gain on perceptual switches is mediated by a flattening of the energy landscape traversed by the network dynamics. Crucially, translating the dynamics of the RNN into an energy-based framework also allowed us to generate a series of predictions that we could later test in functional neuroimaging data.

In recent work[47,71], we have shown that peaks in BOLD within the LC precede large changes in brain state dynamics. Viewed through the lens of dynamical systems theory[72] in which the brain is treated as a dynamical system whose state space (i.e., an instantaneous snap-shot of the activity of all regions of the system) evolves over time shaped by the presence (or absence) of attractors the effect of the LC can be conceptualised as akin to lowering the energy barrier required to escape a fixed-point attractor or as a transient injection of kinetic energy via an external force allowing the brain to reach a novel location in state-space[47]. Crucially, there are two complementary viewpoints from which we can construct an energy landscape; the first allocentric (i.e., third-person view) perspective quantifies the energy associated with each position in state space, whereas the second egocentric (i.e., first person view) perspective quantifies the

energy associated relative changes independent of the direction of movement or the location in state space. The allocentric perspective is straightforwardly comparable to the potential function of a dynamical system but can only be applied to low dimensional data in settings where a position-like quantity is meaningfully defined. The egocentric perspective is analogous to taking the point of view of a single particle in a physical setting and quantifying the energy associated with movement relative to the particle's initial location. An egocentric framework is thus more applicable, when signal magnitude is relative rather than absolute. See materials and methods, and (see **Fig. S4** for an intuitive explanation of the allocentric and egocentric energy landscape analysis on a toy dynamical system).

To characterise the energy landscape traversed by the network dynamics we ran both time-resolved allocentric and egocentric energy landscape analyses. For the allocentric analysis we first had to reduce the dimensionality of the RNN's dynamics by performing a Principal Component Analysis (PCA) on the concatenated activity of the network at gain = 1. The set of PCs was low-dimensional, with 80.58 ± 6.34% of the variance explained by the first principal component (PC_1). Based on this information, we projected the network activity on each trial and for each gain value and timepoint onto the first PC. The resultant low dimensional trajectories all showed a change in direction around the timepoint of the switch in network output from category 1 to category 2 (and v.v.; **Fig. 4A**). This recapitulates a system jumping between attractors, occurring earlier as a function of heightened gain associated with heightened values of γ (**Fig. 4A**)· This switch not only occurred sooner as a function of heightened gain, it also occurred at a higher neural "speed" with the velocity of the trajectory peaking sharply at the point of the switch under high γ, whereas the transition between states was comparatively gradual under low γ (**Fig. 4B**).

With a low dimensional description of our data in hand, we leveraged the relationship between probability and energy in statistical mechanics to construct a measure of the allocentric energy landscape (**Fig. 4C-D**) traversed by the low dimensional dynamics $\left(E_{PC1\tau} = \ln\left(\frac{1}{P(PC1\tau)}\right)\right.$; see methods for derivation) with a window size of τ = 250 ms. Across values of γ, this revealed a potential-like energy landscape with a minimum that evolved with the currently dominant input dimension. To quantify the effect of gain mediated changes on the allocentric energy landscape we devised a measure - neural work - of the "force" exerted on the low dimensional trajectory by the vector field quantified by allocentric energy landscape at each time point in the trial $W_t = -\frac{\mathrm{d}E_t}{\mathrm{dx}} s_t$. Where s_t is the displacement of the PC trajectory in each window, and $\frac{\mathrm{d}E_t}{\mathrm{dx}}$ is the gradient of the energy values computed between the start and end of each window. We found that increasing gain (via increasing γ) increased the magnitude of work done at turning points of the trajectory analogous to the application of an external force (**Fig. 4G**; and equivalent to a change in the dynamical velocity of the landscape, accelerating the change from one perceptual interpretation to another).

Although explanatory useful in understanding the operation of the RNN, the allocentric landscape is not straightforwardly applicable to non-invasive neuroimaging data. In order to compare out network dynamics to neuroimaging data, and with previous work from our group, we inferred an estimate of the egocentric energy landscape (**Fig. 4E-F**) traversed by the dynamics. Specifically, we calculated the mean-squared displacement $MSD_{t,t_0} = \langle \left| \boldsymbol{x}_{t_0+\tau} - \boldsymbol{x}_{t_0} \right|^2 \rangle_n$ of the firing rate of each unit in the RNN in steps of τ = 250 ms, and as we did with the allocentric analysis, calculated the probability - and from this the energy $E_{MSD,\tau} = \ln\left(\frac{1}{P(\mathrm{MSD}_\tau)}\right) -$ associated with each MSD value and time step. In line with our hypothesis, and with previous work from our group, the energy required for large movements in state space (i.e., large MSD values) decreased as a function of γ (**Fig. 4E-F**) analogous to the application of an external force

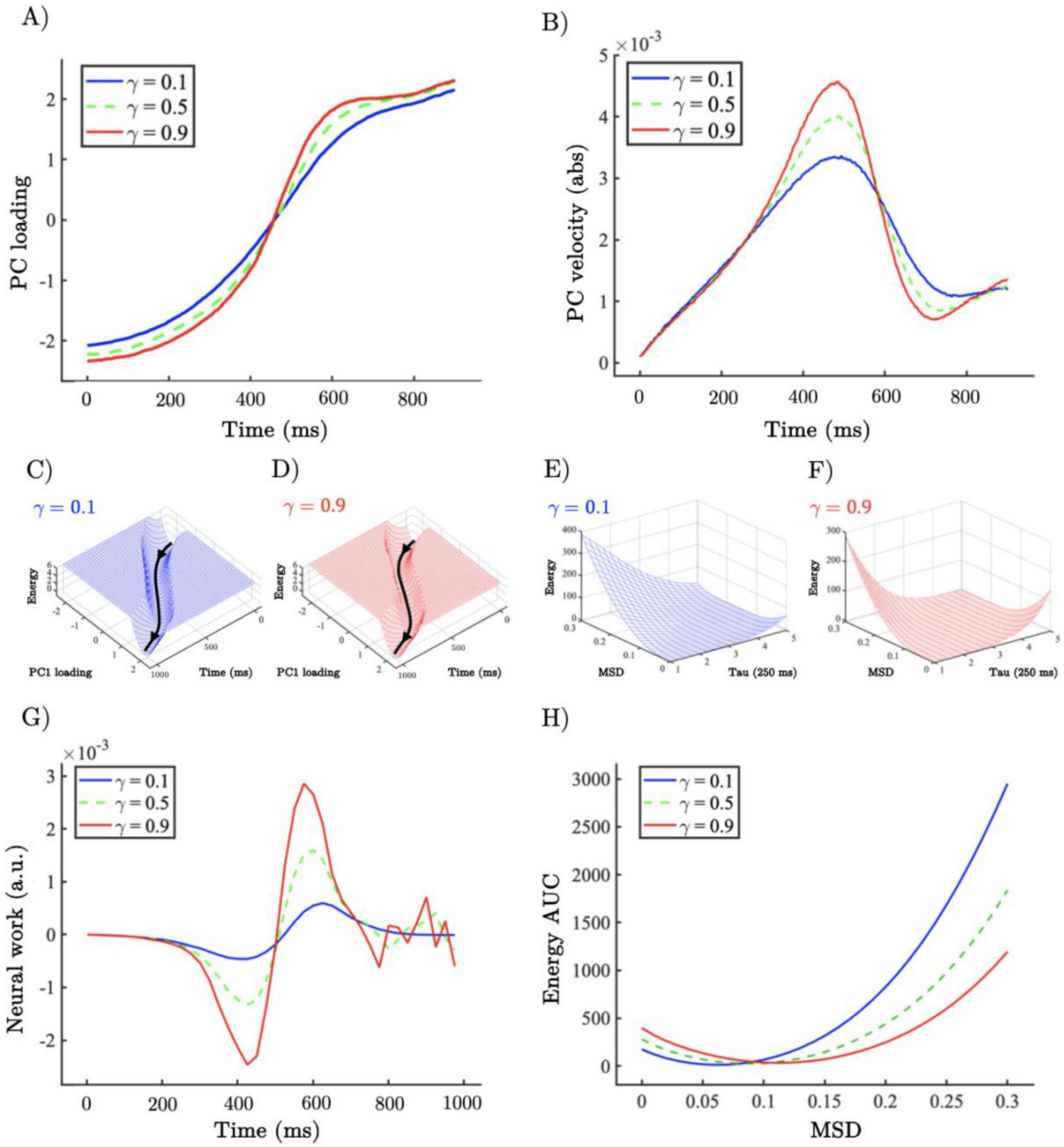


Figure 4

Allocentric and egocentric energy landscape dynamics underlying the perceptual speeding effect of heightened gain.

A) Example network trajectory projected onto PC1 and averaged across trials for low (0.1; solid blue), medium (0.5; dotted green), and high (0.9; solid red) *γ* for the $u_1 \rightarrow u_2$ condition. B) (abs) Velocity of PC1 trajectories across low (0.1), medium (0.5), and high (0.9) *γ*. C-D) Allocentric landscapes for low (0.1; blue) and high (0.9; red) *γ* conditions. Trial averaged PC1 trajectory shown in black. For purposes of visualisation energy values > 6 are set to a constant value. E-F) Egocentric landscapes for low (0.1; blue) and high (0.9; red) *γ* conditions. G) (Allocentric) neural work for low (0.1), medium (0.5), and high (0.9) *γ*, averaged across networks and conditions. H) Egocentric AUC for low (0.1), medium (0.5), and high (0.9) *γ*, averaged across networks and conditions.

transiently increasing the kinetic energy of a particle. To quantify the degree of flattening we calculated the area under the curve across values of γ showing a substantial reduction in the energy associated with large MSD values as a function of heightened γ (and therefore gain; **Fig. 4H**).

These results reinforce our previous work and clearly demonstrates that the implementation of neuromodulatory-mediated dynamics in the RNN acted in a similar fashion to previously observed patterns in resting-state fMRI[26]. In addition, our results confirm that the putative impact of the release of noradrenaline from the locus coeruleus can change the manner in which brain states evolve over time facilitating the navigation of otherwise difficult state transitions[26].

The low-dimensional signature of ambiguity resolution and perceptual change

Having confirmed our hypothesis about the speeding effect of gain in our RNN model we next sought to test the predictions in the human brain - i.e., examining whether the increase in neural speed and the flattening of the energy landscape observed in the RNN were also present in functional neuroimaging data. To this end, we re-analysed an existing BOLD dataset collected while participants performed a similar version of the ambiguous figures task to identify the lowdimensional patterns that occurs during the perceptual change.

We were, however, left with a dilemma: RNNs provide a proof-in-principle of how computations can be instantiated in neural networks, however there are key differences between artificial neural networks and the human brain that require careful consideration[73,74]. While both RNNs and the brain are thought to compute through dynamics [75], the human brain is comprised of highly specialised neural circuits that have been shaped over evolutionary time to perform a range of highly idiosyncratic functions that matter for adaptive behaviour[75], but aren't necessarily related to task-switching. So where in the brain should we look for the same lowdimensional signatures we observed in the RNN as a function of gain? Rather than select a particular region *a priori*, we instead opted for a data-driven approach - principal components analysis (PCA) - which summarizes regional timeseries concatenated across all subjects and trials into a set of low-dimensional patterns that can then be interrogated in a similar fashion to the activity of the RNN (see Methods for details). Consistent with previous work[76], a small number of principal components (PCs) mapped onto distributed regions across the brain (**Fig. 5A**) and explained a substantial proportion of the variance observed in the task ($PC_{1–3}$ explained 32% of the total variance).

To isolate the low-dimensional component that best reflected the task (**Fig. 1A**), we performed a principal component regression[77] that modelled the switch point of each trial using the loadings of the top 3 PCs calculated from fMRI data. PC_1 was not selectively aligned with switches, both PC_2 and PC_3 showed a pronounced, isolated peak around the switch point across trials (**Fig. 5B**), with PC_2 showing the most robust task-related engagement (**Fig. 5B** & Fig. S5). To ensure that these results could not be explained by the spatial autocorrelation inherent within the PC maps, we created a null distribution of regression coefficients calculated using the same statistical model but with block-resampling applied to the switch times in the design matrix. The dotted grey line in **Fig. 5B** denotes the 95^{th} percentile of the null distribution, and clearly shows that the engagement of both PC_2 and PC_3 during the switch point was greater than to be expected by chance. Furthermore, to validate that the perceptual switch was predominantly represented by PC_2 and PC_3 (**Fig. 5D**), we conducted a regression with these two PCs as predictors and the evoked activity derived from the original BOLD time-series as the dependent variable. The resulting variance accounted for was 88% (R^2 = 0.88, β = 0.99, $p = 9.2\times10^{-178}$).

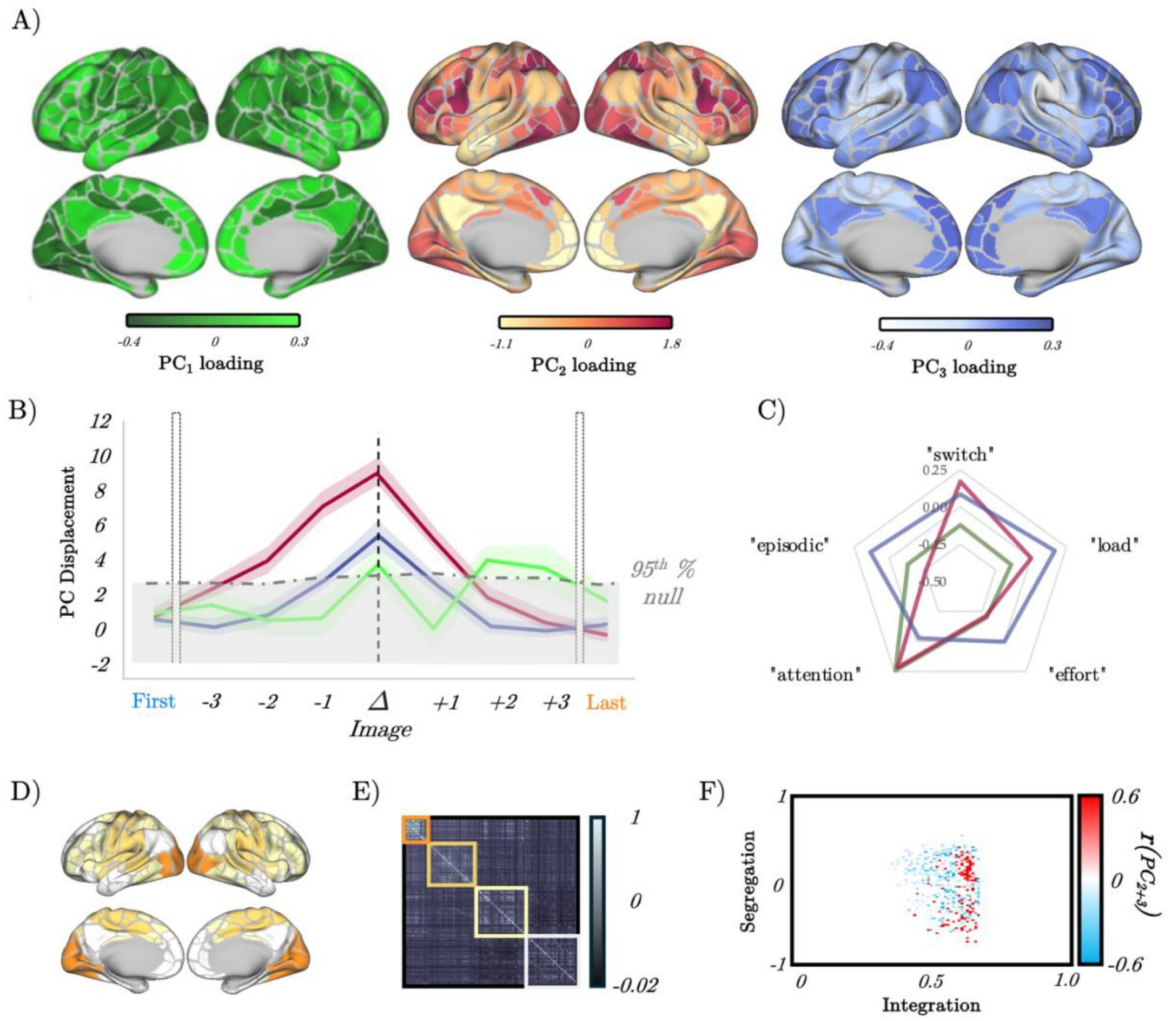


Figure 5

Low-dimensional switch-related dynamics and connectivity.

A) spatial loadings of PC_1 (green), PC_2 (red) and PC_3 (blue); B) Mean absolute β loading (solid lines) and group standard error (shaded) of PC_1 (green), PC_2 (red) and PC_3 (blue), organized around the image switch point (Δ) - the dotted grey lines show the 95^{th} percentile of the null distribution of a block-resampling permutation; C) radar plot showing the partial correlations of PC_1 (green), PC_2 (red) and PC_3 (blue); D) Evoked Brain activity of PC_2 + PC_3 during the perceptual switch. E) Group averaged functional connectivity and module assignments using a Louvain analysis - three clusters were observed. F) Pearson's correlation between the sum of PC_2 and PC_3 (per subject) and a joint-histogram comparing Integration (participation coefficient) and Segregation (module-degree Z-score); $p < 0.05$ following permutation testing.

To determine whether PC_2 or PC_3 was a better index of perceptual switching, we then correlated the spatial loadings of PC_2 and PC_3 with the spatial map associated with the term "switching" from a meta-analysis performed on the *neurosynth* database[78]. We observed a significant positive correlation between the map for "switching" and both PC_2 ($r = 0.453$, $p = 3.041\times10^{-18}$), and PC_3 ($r = 0.115$, $p = 0.037$), however the correlation for PC_3 was much lower than PC_3, suggesting that PC_2 was a better match for "switching". The spatial map of PC2 was also positively correlated with other terms putatively associated with the ambiguous figures task (notably, "effort", "load" and "attention"; all $r > 0.2$; and not with "episodic", which was included as a negative control), a partial correlation analysis revealed that PC_2 was selectively associated with "switching" and "attention" (**Fig. 5C**). Given the multifaceted nature of the ambiguous figures task, the convergence between brain maps for "switching", "attention", and "effort" was to be expected, and we therefore did not try to dissociate them in further analysis.

Before turning to the predictions of the RNN we first sought to validate the face validity of focusing on a limited number of principle components. In previous work, we have linked the impacts of NA on systems-level neural dynamics to alterations in network topology[47,79,80], with NA increasing large-scale network integration. Given that PCA naturally captures patterns of covariance between regions, we expected to see that the observed time signatures of PC engagement at the switch-point should coincide with similar measures of network integration. To test this hypothesis, we clustered the time-averaged functional connectivity matrix using a hierarchical modular decomposition approach (see Methods) - doing so revealed three main clusters (**Fig. 5E**). For each participant, we used this matrix and the three clusters to estimate the amount of integration (using the participation coefficient) and segregation (using the moduledegree Z-score, see Methods) of each region. We then correlated a joint histogram of these measures with the sum of subject-specific regression coefficients for PC_2 and PC_3 and observed a robust correlation with integration (**Fig. 5F**; $p < 0.05$ following permutation testing). These results clearly demonstrate the highly convergent nature of PCA and our previous network-based approaches.

Confirmation of model predictions in whole-brain BOLD data

Based on the patterns observed in the RNN (i.e., those in **Fig. 2**–**4**), we hypothesized that the energy landscape topography would decrease, and the velocity of the low dimensional brain patterns would peak at the switch point. Given the prominent role in switching, we focused our analysis on the PC_2 time series. To estimate the (egocentric) energy landscape, we first estimated the mean displacement of PC_2 by averaging the β value around the switch-point and then divided this term by the logarithm of the inverse probability of the loading of PC_2, which was also inferred from the GLM. Using this approach, we observed that PC_2 was maximally displaced at the perceptual change, suggesting that the brain state showed a substantial shift from baseline during the perceptual change. Energy ($\log[1/p_{switch}]$; see Methods) showed a U-shaped pattern around the perceptual change point - i.e., with a minimum value in the perceptual change along with the first and last images (**Fig. 6B-D**). To relate this measure to the energy landscape framework, and to control by the specific displacement occurring at each image, we then calculated the ratio between energy and the mean displacement (i.e., energy landscape 'depth'; **Fig. 6A**). As predicted, the brain-state reduced the amount of energy per displacement towards its minimum around the perceptual change (**Fig. 6B**). We interpret this set of results as the system flattening the energy landscape, reducing the energy (i.e., higher system changes become more common) required for large displacement values effectively generating a 'network reset'[34,36,81] of the brain state, which ultimately facilitated an updating of the content of perception.

To analyse speed-evoked changes in brain trajectories, we used a GLM to analyse each PC time series as a function of each perceptual switch. Our design matrix included the first and last images seen in each set, along with the three images leading up to the switch, the switch trial itself and the three images following the switch (see Methods for details). This approach thus allowed us to track the low-dimensional signature of the brain through the processing and resolution of perceptual

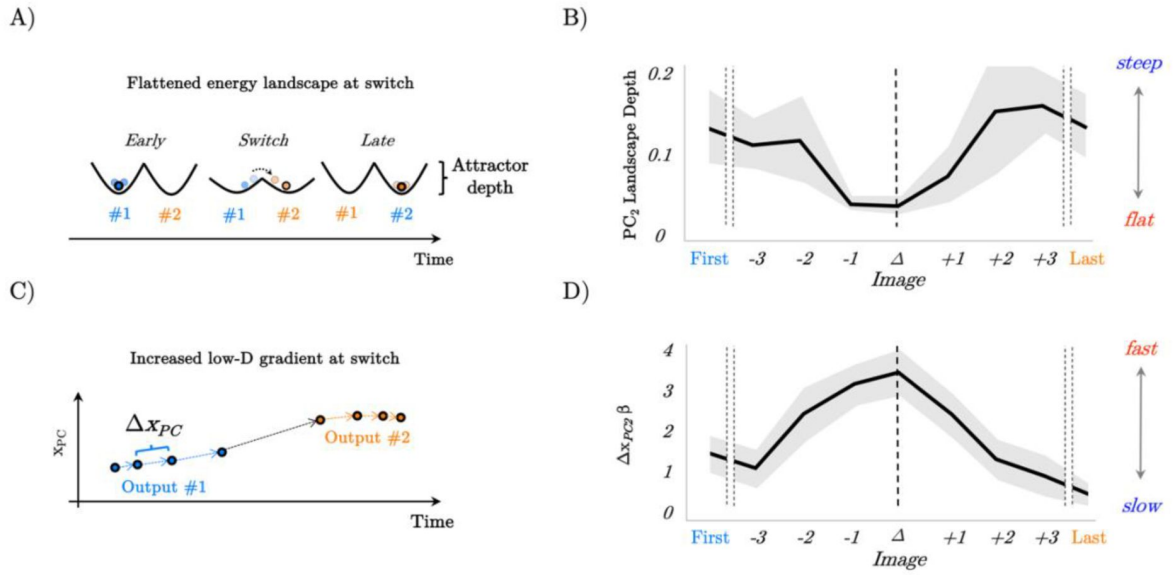


Figure 6

Confirmation of model predictions in whole-brain BOLD data.

A) analysis of the RNN also predicted that the energy landscape dictating the likelihood of state transitions should be flat (i.e., have a small attractor depth) at the switch point; B) the energy landscape was demonstratively flatter (quantified as surprisal over brain activity displacement) at the switch-point; C) by interrogating the low-dimensional trajectories in the RNN, we predicted that there should be a peak in the gradient of the loadings in principal component space at the switch point between output #1 and output #2 ; D) the gradient (Δx_{PC}) of the β loading of PC_2 as a function of the switch point.

ambiguity. As predicted (**Fig. 6C**), we found evidence that PC_2 showed a peak in velocity at the change point (**Fig. 6D**) providing confirmatory evidence that the low-dimensional brain state dynamics observed in whole-brain fMRI were highly similar to those observed in the trained RNN.

Discussion

Here, we studied the relationship between the ascending arousal system, low-dimensional neuronal trajectories and energy landscape dynamics during a perceptual switch task. Our results provide evidence that the ascending arousal system is involved in the modulation of dynamic brain state topography during task-relevant perceptual switches. We found that pupil diameter tracked with ambiguity of task stimuli and was directly related to the speed of perceptual switches (**Fig. 1**). Next, we confirmed that this process could be replicated in an RNN Model (**Fig. 2**) of perceptual change detection where the gain of the activation function was updated dynamically by the uncertainty of the network's classification output (**Fig. 2**–**3**). We then used this model to generate two key predictions: around the time of the perceptual switch brain state velocity should peak, and the egocentric energy landscape should flatten which we confirmed in neuroimaging data (**Fig. 4**–**6**). Together, these results suggests that the ascending arousal system facilitates state changes in the content of perception by transiently increasing neural gain - acting in a manner analogous to an external forcing function transiently increasing kinetic energy in the system - flattening the ego-centric energy landscape and thereby reducing the energy needed to reset the system topography in an adaptive and task dependent manner.

The relationship between perception and pupil diameter found here is consistent with the role of the ascending neuromodulation in cognition and attention[61,65]. For instance, the LC dynamically changes its activity according to external and cognitive demands imposed on the system[51,61,65,82,83]. Importantly, our results extend these findings by suggesting a more precise role for LC-mediated alterations in neural gain. Specifically based on the pupil dynamics in our task and previous experimental and theoretical work, we hypothesised that neural gain should change dynamically as a function of uncertainty (operationalised here as perceptual ambiguity) via the recruitment of the LC (along with other structures in the ascending arousal system), which then subsequently increases brain-wide communication by increasing the gain in targeted brain regions[11,16,49,83]. In the pupillometry data pupil diameter (which is an indirect marker of the noradrenergic system[32,37]) increased as a function of perceptual ambiguity, which rose sharply in the few images prior to the reported perceptual change (**Fig. 1D**). Based upon this finding we then implemented an analogous mechanism in our pretrained RNN by making gain depend upon the entropy of the network's classification which acted as a forcing function transiently increasing gain when the input became ambiguous, which, in line with our hypothesis lead to earlier perceptual switches. We chose to use an RNN, instead of a simpler (more transparent) model as we wanted to use the RNN as a means of both hypothesis generation and hypothesis testing. Specifically, unlike more standard neuronal models which are handcrafted to reproduce a specific effect, when building an RNN the modeller only specifies the network inputs, labels, and the parameter constraints (e.g. Dale's law) in advance. The dynamics of the RNN are entirely determined by optimisation. Post-training manipulations of the RNN are not built in, or in any way guaranteed to work, making them more analogous to experimental manipulations of an approximately task-optimal brain-like system. Confirmatory results are arguably, therefore, a first steps towards an *in vitro* experimental test.

Thus, we provide early empirical and computational evidence that ascending neuromodulatory activity facilitates state changes in perception under conditions of perceptual ambiguity[31,45,46] when a stimulus is task relevant. Importantly, we do not expect that our results will generalise to experimental setting when a stimulus is not task relevant. We can make sense of this computationally by imagining the gain dynamics in our model if we added in a second taskirrelevant condition where at the beginning of each trial the model was given a cue

indicating whether it would have to simply “maintain fixation” or readout the category of the input. In the presence of the task irrelevant cue the model would readout the “maintain fixation” action with high certainty and thus not ramp up gain. We hypothesise therefore that the pupil dynamics observed in the task will depend on participants task-set. Indeed, there is evidence from a recent multistable perception experiment showing that arousal-related changes in pupil dilation disappear when the stimulus is not task-relevant. The authors of the study attribute the arousaldependent pupil dilation to task-execution. This explanation, however, could not explain the ramping of pupil diameter in our task where the participants perform an action on every trial. Instead, based upon the workings of our computational model, we hypothesise that arousalbased changed in pupil diameter are driven by task-set related uncertainty and thus will depend on task-relevance rather than task-execution per se.

A core neuroanatomical property of the LC noradrenergic system is that a relatively small number of neurons (~fifty thousand in an adult human) projects to almost all brain regions[38,84]. This organisation implies that the LC acts as a low dimensional modulator of the much more high-dimensional cerebral cortex. Subtle changes in the activity of LC can have significant effects on how different brain regions communicate[49,65,82,85–87]. The mechanism of gain modulation in our model was, likewise, dependent of a low-dimensional process with the network output altering the gain uniformly across the full network. At a neuronal level NA increases excitability by liberating intracellular calcium and opening (or closing) voltage-gated ions channels[11,65]. In our model this global increase in excitability increased the speed of perceptual switches by recruiting inhibitory units to more rapidly actively inhibit the population encoding the initially dominant stimulus. At a population level the interaction between excitatory and inhibitory units led to the emergence of a gain-dependent oscillatory regime which suppresses the currently active population encoding the initially dominant stimulus and boosts the competing quiescent population. At the scale of the full network the gain mediated changes resemble the transient application of an external forcing function pushing the network trajectory in the direction of the new percept which, from the perspective of the allocentric landscape, manifests as a spike in neural work at turning points in the network’s low-dimensional trajectory leading up to and following the perceptual switch. From the egocentric perspective this is characterised by a flattening of the landscape analogous to an externally driven increase in kinetic energy making large changes in the location of a particle more likely.

In line with the predictions of the RNN in our analysis of the BOLD data, we showed that the velocity of the low dimensional brain state trajectory most associated with perceptual switching increased significantly during the point of reported perceptual change in comparison (**Fig. 5B**), which we interpret as the brain moving from one attractor to another (**Fig. 6A**). Importantly, we showed that around the perceptual switch, the energy needed for each unit of change in brain state (i.e., displacement) is smaller than at other points in the task (**Fig. 6A-B**). Under the (egocentric) energy landscape framework[47,60], this tells us that the landscape is flattened, and the energy required to transition between states is reduced. Together with the pupillary findings (**Fig. 1**), the computation model (**Fig. 2**–**4**), and replication from former results[47,60] (**Fig. 5E-F**), we propose that the ascending neuromodulatory system is responsible for the large-scale flattening of the egocentric energy landscape which facilitating changes in task-relevant perceptual content.

This work is not without limitations. First, the pupil diameter dataset and the fMRI analysis came from different participants, such that the link between the pupil diameter and the fMRI results is inherently indirect. Moreover, differences in task timing, structure, and instructions between the fMRI and pupil experiments add complexity to interpreting the results. For instance, the fMRI task includes jittered inter-trial intervals (ITIs) and catch trials, features absent in the pupil task, which presents a more rapid stimulus sequence. These differences may have influenced perceptual switch points and task behavior across experiments. Additionally, the specificity of the pupil diameter as a marker of the LC activity is under active debate[37]. For instance, there is evidence

suggesting a role of the superior colliculus, the dorsal raphe nucleus and central cholinergic system in driving pupil dilations[43,75,76]. Although there is uncertainty regarding whether these other nuclei are directly related to pupil dilation, or only indirectly via their connections with other neural regions and nuclei. Despite this, we believe that our pupillometry dataset captures an important function of the noradrenergic system in cases of task-relevant perceptual ambiguity as there is strong evidence showing that pupil diameter is a reliable marker of noradrenergic activity during evoked cognitive tasks[44,85,89,91]. Additionally, the sample size of our fMRI study makes it difficult to generalize our results. In spite of this, the converging evidence from the pupillometry dataset, the fMRI dataset and the computational model, supports the role of the ascending neuromodulation in mediating task-relevant perceptual switches. Future work is needed both in humans, with higher sample sizes utilizing fMRI and eye-tracking recordings, as well as animal studies, to directly modulate and record the LC activity in a task manipulating perceptual uncertainty.

Conclusion

In summary, we provide computational and empirical evidence for the association between neuromodulation, pupil dilation, and (egocentric) energy landscape flattening in task-relevant perceptual switches. Our results strengthen our understanding of the neurobiological processes underpinning moment-by-moment adaptive changes to perception. Specifically, we suggest that the widespread excitatory projections of the noradrenergic arousal system mediate the systems-level reconfigurations of cortical network architecture[84,85] via uncertainty driven alterations in neural gain. This suggests that more highly conserved features of the nervous system may play a role in driving task-relevant switches in the contents of perception

Methods

Overview of empirical data

There were two independent groups analysed in this study: 35 subjects performed a perceptual decision-making perceptual task while pupil diameter was recorded; and a separate group of 17 subjects performed a version of the task adapted for the MRI scanner.

Perceptual Task

Twenty picture sets were used in which line drawings of common objects morphed over 15 iterations into a different object (**Fig. 1A**). Picture sets were selected from a larger set validated in an earlier study[48]. In the original study, participants reported verbally what they saw by typing in the name of the object. This reporting method guaranteed that participants could freely indicate what they saw without being restricted by categories (e.g., forced choice). Picture sets for the current study were selected with the criterion that all sets were perceived categorically in the normative study (i.e., that the majority of participants in the normative study categorized each picture they saw as either the first object or second object in the set[17]). Selecting only the categorically perceived image sets guaranteed that pictures in the middle of the morphing sequence were not simply 'noisier' than pictures at the beginning or end. In other words, the ambiguous images were still easily categorised by participants as either object 1 or object 2. All images were a standard size (316×316 pixels) and were displayed on a white background. In addition, in the fMRI study, participants were presented with two kinds of control picture sets to ensure that they were responding to changes in the pictures in the set rather than simply to the position in the set (e.g., always switching after the 8th picture). In these control picture sets, a salient deviating picture was presented either after three pictures or after thirteen pictures resulting in an early or late abrupt shift. Those sets served as controls and were not analysed further.

The picture morphing task consisted of five experimental runs. We randomized the order in which the picture sets were presented in each run and kept this randomized order consistent across participants. Picture morphing in each picture set occurred over fifteen discrete steps, each corresponding with the acquisition of a whole-brain image. In the fMRI experiment, each picture within a set was presented for two seconds. Pictures were randomly intermixed with eight interstimulus-intervals (2, 4, 6 or 8 seconds) during which participants saw a fixation cross. In the eyetracking experiments, each picture was presented for 500ms, followed by a fixation cross of 2 seconds. Participants provided their responses in the scanner using two buttons on a four button Cedrus fibre optic system. In a two-alternative forced-choice task, participants were asked to press the first button when they 'saw the first object' and the second button when they 'saw the second object' - this ensured that there was not a motor confound present on only the switch trials. All participants were ignorant as to the identity of the second object in each picture set. At the end of each set of 15 images the word END was presented for 2 s to indicate that the next picture set would begin shortly. Participants provided their responses in the fMRI scanner using a Cedrus fiber-optic response system with four buttons. For the two-alternative forced-choice task, participants were instructed to press the first button when they 'saw the first object' and the second button when they 'saw the second object.' This design ensured that motor responses were not confounded with perceptual switches, as responses occurred on both switch and nonswitch trials. Importantly, participants were not informed about the identity of the second object in each picture set beforehand. At the end of each sequence of 15 images, the word 'END' was displayed for 2 seconds to signal the conclusion of that picture set and the imminent start of the next one.

Participants

A total of seventeen (6 male) neurologically healthy participants with normal or corrected to normal vision took part in the fMRI study (mean age 27.65 ± 8.01). Fifteen were right-hand dominant. A separate cohort of 35 participants performed the task while simultaneous pupil diameter was recorded using an eye tracker device (SR Research, 1000 Hz). None of the participants had a history of brain injury. Participants received $30 for their participation. All participants provided informed consent prior to participation. The research protocol was approved by the Office of Research Ethics at the University of Waterloo and the Tri-Hospital Research Ethics Board of the Region of Waterloo in Ontario, Canada.

Pupillometry

Fluctuations in pupil diameter of the left eye were collected using an Eyelink 1000 (SR Research Ltd., Mississauga, Ontario, Canada), with a 1 kHz sampling frequency. Blinks, artifacts, and outliers were removed and linearly interpolated[92]. High-frequency noise was smoothed using a second-order 2.5-Hz low-pass Butterworth filter. To obtain the pupil diameter average profile, data from each participant were normalized across each trial (corresponding to the 15 consecutive image set). This allowed us to correct for low-frequency baseline changes without eliminating the load effect and baseline differences due to load manipulations[93,94].

Recurrent Neural Network Modelling

We used PyTorch[95] to implement and train 50 continuous-time recurrent neural networks that we constrained to respect Dale's law (N_{E+I} = 40, 80% excitatory N_E = 32, and 20% inhibitory N_I = 8) using the procedure set out in[63]. The dynamics of each network evolved according to the following system of stochastic differential equations:

$$dx = \frac{1}{\tau}(-x(t) + W^{rec}r(t) + W^{in}u(t))dt + dW$$

Where $x \in \mathbb{R}^{N\times 1}$ represents the sub-threshold activation of each unit, $u \in \mathbb{R}^{2\times 1}$ the external input into the network, $W^{rec} \in \mathbb{R}^{40\times 40}$ the recurrent weights, $W^{in} \in \mathbb{R}^{40\times 2}$ the input weights, and τ the time constant which we set to 100ms. In addition to task input each unit in the network was driven by a Weiner process dW. The subthreshold activation variable x was converted into a vector of instantaneous firing rates by applying a sigmoid function $r = \frac{1}{1+e^{(-gx)}}$ where $g \in \mathbb{R}^{N\times 1}$ is a vector containing the gain control parameter of each unit's activation function that was multiplied element wise with x. Network outputs $z \in \mathbb{R}^{2\times 1}$ were given by a linear readout of the excitatory population's firing rate $z = W^{out}r_E$. Where $W^{out} \in \mathbb{R}^{N_E\times 2}$. The network's choice at each time point was the maximum of the two-dimensional output z.

We imposed Dale's law on the recurrent weights of the network by parametrising the weight matrix with a mask $W^{mask} \in \mathbb{R}^{40\times 40}$ which contained zeros in the leading diagonal (removing self-connections), +1 in all non-diagonal entries of the first 32 rows/columns and -1 in the remaining 8 rows/columns. We obtained the constrained recurrent weight matrix by multiplying the absolute value of the trained weights element wise with the mask $W^{rec} = \left|W^{rec}_{plastic}\right| \odot W^{mask}$ thereby imposing an 80/20 E/I ratio. Similarly, we constrained the projection of the input to the network and the readout projection to be strictly positive by taking the absolute value of the trained input and output weights $W^{in} = \left|W^{in}_{plastic}\right|$, $W^{out} = \left|W^{out}_{plastic}\right|$.

Following standard practice[63], we simulated the network by discretising the system using a Euler-Maruyama integration scheme where $\alpha = \frac{dt}{\tau}$, and $\sigma_{rec} = 0.01$.

$$x(t+\Delta t) = (1-\alpha)x(t) + \alpha\left(W^{rec}r(t) + W^{in}u(t)\right) + \sigma_{rec}\sqrt{\Delta t}\,\mathrm{N}(0,1)$$

Each network was trained by optimising $W^{in}_{plastic}$, $W^{rec}_{plastic}$, and $W^{out}_{plastic}$, to minimise a cross entropy loss function through 1000 iterations of back propagation through time[96] with ADAM[97]. Batches consisted of single trials which for our simple task (described below) was sufficient for each network to converge on near perfect behavioural accuracy. All training was performed with the gain control parameter set to 1. Again following standard practice, we trained the networks with a relatively large time step of Δt = 200 ms. Following training, to ensure numerical stability we exported the trained weights into MATLAB and simulated the system with a bespoke numerical integration scheme with Δt set 1ms.

The task consisted of a simple change detection paradigm analogous to the task performed by our human participants. Specifically, at each time point the network was fed a two-dimensional input $u(t) = [u_i\ u_2]^T$ with each column representing the "sensory evidence" for each of the two stimulus categories. The task lasted for 1 second of simulation time beginning with maximum evidence for one of the two categories $u(t) = [1\ 0]^T$ and over the course of each trial changed linearly such so that at the half way point of the simulation the sensory evidence for each stimulus category changed was perfectly matched category $u(t) = [.5\ .5]^T$ and by the final time-step consisted of maximum evidence for the second stimulus category $u(t) = [0\ 1]^T$. We trained the network to output a response for stimulus category 1 whenever $u_1 > 0.5$, and $u_2 < 0.5$, and category 2 whenever $u_1 < 0.5$, and $u_2 > 0.5$.

To test our hypothesis that perceptual uncertainty increases neuromodulatory via phasic bursts in the noradrenergic locus coeruleus we made gain time dependent with dynamics governed by a linear ODE with a forcing term proportional to the uncertainty (i.e. the entropy $H(z) = \Sigma_i p(z)_i \ln(p(z)_i)$) of the network's readout.

$$dg = \frac{1}{\tau}\left(\left(g_{tonic} - g(t)\right) + \gamma H(z)\right)dt$$

Where $p(z)$ is obtained by passing $z(t)$ through a softmax function at each time step of the simulation $p(z)_i = \frac{\exp(\omega z_i)}{\sum_j^K \exp(\omega z_j)}$ with inverse temperature parameter $\omega = 0.25$. When the network approaches the half way point in the trial input is maximally ambiguous and the distribution $p(z)$ approaches a uniform distribution leading $H(z)$ to approach its maximum value which in turn leads to a phasic increase in gain (with magnitude γ). In the absence of forcing (i.e. under conditions of perceptual certainty) gain decays exponentially to its tonic value ($g_{tonic} = 1$).

To study how the population dynamics of the trained networks changed as a function of gain in a shared space we performed a Principal Component Analysis (PCA) on the concatenated activity of the network at $\gamma = 0$. The set of principal components was highly low-dimensional, with 80.58 ± 6.34% of the variance explained by the first principal component (PC1). We then projected the trial averaged activity at each gain value at each timepoint onto the top PC.

Energy Landscape Analysis

Leveraging previous work from our group[47] we constructed a measure of the energy landscape traversed by each network through an analogy to the relationship between probability and energy in statistical mechanics[98] given by the Boltzmann distribution.

$$p_i = \frac{1}{z}e^{-\beta E_i}$$

Where p_i denotes the probability of each state, E_i the energy of each state, β the thermodynamic beta, and z the canonical partition function. Solving for E_i we obtain:

$$E_i = \frac{1}{\beta}\ln\left(\frac{1}{zp_i}\right)$$

Instead of inferring the probability distribution from the energy of a state as in done in physics we used the fitdist function in MATLAB with a Gaussian kernel $\left(P(x) = \frac{1}{4n}\sum_{i=1}^{n} K\left(\frac{x}{4}\right)\right.$, where $\left.K(u) = \frac{1}{2\sqrt{\pi}}e^{-\frac{1}{2}u^2}\right)$ to infer the probability of the state, and then solved for the energy. As $\int_{-\infty}^{+\infty} P(x)dx = 1$ by construction the partition function z, which we define here to be the integral of the pdf, is equal to 1, which after setting $\beta = 1$ yields:

$$E_i = \ln\left(\frac{1}{p_i}\right)$$

For the allocentric landscape analysis we defined the state of the system in terms of the trial averaged loadings on PC1 which we divided into 250 ms windows. For the egocentric landscape analysis, we calculated the mean-squared displacement (MSD) of the activity of the RNN at each time point τ_0 relative to reference point $\tau_0 + \tau$.

$$MSD_{\tau,\tau_0} = \langle \left| \boldsymbol{x}_{\tau_0+\tau} - \boldsymbol{x}_{\tau_0} \right|^2 \rangle_n$$

For congruency with the allocentric analysis we increased τ and τ_0 in steps of 250 ms starting 1s into the trial and ending with a maximum difference between τ and τ_0 of 5 s to ensure that all steps had equivalent window sizes.

Following the physical analogy, we think of the state of the system, PC1 loadings in the allocentric analysis, and *MSD* in the egocentric analysis, as akin to the location and movement of a particle respectively. Positions in state space with low energy have a higher probability of being occupied, and systems with a higher average energy have a more uniform probability distribution making large jumps in the position of a particle more likely (i.e. lower energy for large MSD values). See supplementary material (**Fig. S4**).

To quantify the effect of gain mediated alterations to the topography of the allocentric energy landscape we devised a novel measure - neural work - of the force (which in classical mechanics is equal to the negative gradient of potential energy) exerted on the low dimensional neural trajectory by the vector field quantified by the allocentric energy landscape at each time point in the trial.

$$W_t = -\frac{\mathrm{d}E_t}{\mathrm{dx}} s_t$$

Where s_t is the displacement of the PC trajectory, and $\frac{\mathrm{d}E_t}{\mathrm{dx}}$ the energy gradient. We computed s_t from the (absolute) difference between PC1 loadings at the start and end of each time window, and $\frac{\mathrm{d}E_t}{\mathrm{dx}}$ from gradient of energy values at the start and end of each time window.

MRI Data

Functional data were acquired using gradient echo-planar T2*-weighted images collected on a 1.5 T Phillips scanner located at Grand River Hospital in Waterloo, Ontario (TR = 2000 ms; TE = 40 ms; slice thickness = 5 mm with no gap; 26 slices/volume; FOV = 220×220 mm^2; voxel size = 2.75 × 2.75 × 5mm^3; flip angle = 90°). Each experimental run consisted of 26 slices per volume and 285 volumes. At the beginning of each run, a whole brain T1-weighted anatomical image was collected for each participant (TR = 7.4 ms; TE = 3.4 ms; voxel size = 1×1×1mm^3; FOV = 240×240 mm^2; 150 slices with no gap; flip angle = 8°). The experimental protocol was programmed using E-Prime experimental presentation software (v1.1 SP3; Psychology Software Tools, Pittsburgh, PA). Stimuli were presented on an Avotec Silent Vision™ fibre-optic presentation system using binocular projection glasses (Model SV-7021). The onset of each trial was synchronized with the onset of data collection for the appropriate functional volume using trigger pulses from the scanner.

fMRI Data Preprocessing

After realignment (using FSL's MCFLIRT), we used FEAT to unwarp the EPI images in the y-direction with a 10% signal loss threshold and an effective echo spacing of 0.333. Following noisecleaning with FIX (custom training set for scanner, threshold 20, including regression of estimated motion parameters), the unwrapped EPI images were then smoothed at 6-mm FWHM,

and nonlinearly co-registered with the anatomical T1 to 2-mm isotropic MNI space. Temporal artifacts were identified in each dataset by calculating framewise displacement (FD) from the derivatives of the six rigid-body realignment parameters estimated during standard volume realignment[99], as well as the root mean square change in BOLD signal from volume to volume (DVARS). Frames associated with FD > 0.25 mm or DVARS > 2.5% were identified; however, as no participants were identified with greater than 10% of the resting time points exceeding these values, no trials were excluded from further analysis. There were no differences in head motion parameters between the five runs ($p > 0.500$). Following artifact detection, nuisance covariates associated with the six linear head movement parameters (and their temporal derivatives), DVARS, physiological regressors (created using the RETROICOR method), and anatomical masks from the cerebrospinal fluid and deep cerebral white matter were regressed from the data using the CompCor strategy[100]. Finally, in keeping with previous time-resolved connectivity experiments[101], a temporal band pass filter ($0.0071 < f < 0.125$ Hz) was applied to the data.

Brain Parcellation

Following preprocessing, the mean time series was extracted from 375 predefined regions of interest (ROIs). To ensure whole-brain coverage, we extracted the following: (a) 333 cortical parcels (161 and 162 regions from the left and right hemispheres, respectively) using the Gordon atlas[102]; (b) 14 subcortical regions from the Harvard-Oxford subcortical atlas (bilateral thalamus, caudate, putamen, ventral striatum, globus pallidus, amygdala, and hippocampus; *https://fsl.fmrib.ox.ac.uk/*); (c) 28 cerebellar regions from the SUIT atlas[103] for each participant in the study.

Neuroimaging Analysis

In order to analyse task evoked activity related to stimulus presentations, we first performed a principal component analysis (PCA)[76] on the pre-processed BOLD time-series (per subject/session), to extract orthogonal low-dimensional time-series. The top 3 PCs explained ~30.6% of the variance. The time-series of these PCs was entered into a general linear model, in which we modelled the following 9 event-types across an entire session, centred around the perceptual switch point, which changed on a trial-by-trial basis: the first two images (modelled as a single regressor), the seven images surrounding each perceptual change (i.e., the switch trial and the three images surrounding the change point, modelled as seven separate regressors) and the last two images (modelled as a single regressor). Each of the event onset times was also convolved with a canonical hemodynamic response function. This left us with nine unique β values per principal component, which we could use to determine how each PC differentially engaged as a function of the task. To test the hypothesis that the rate of change of PC engagement peaked at the perceptual change point, we calculated the difference between the β value for each of the top 3 PCs for each of the 9 event-types, and then plotted the resultant series in order to identify whether a peak occurred at the perceptual switch point (i.e., the middle β value in the series). A block-resampling null (n = 5,000 permutations) was used as a permutation test ($p < 0.05$).

Spatial maps associated with the terms: “switching”, “effort”, “attention”, “perception” and “load” were downloaded from the *neurosynth* repository[78] and mapped into our parcellation space by calculating the mean value within each independent parcel. These values were then correlated with the spatial loading of each of the top 3 PCs. A separate partial correlation analysis was conducted in which the same correlation was estimated after controlling for each of the other spatial maps.

Topological analyses

A hierarchical modularity approach was used to collapse the mean time-averaged correlation matrix across participants into a set of four spatially non-overlapping modules. Briefly, this involved running the Louvain modularity algorithm, which iteratively maximizes the modularity

statistic, Q, for different community assignments until the maximum possible score of Q has been obtained.

$$Q_T = \frac{1}{v^+}\sum_{ij}\left(w_{ij}^+ - e_{ij}^+\right)\delta_{M_iM_j} - \frac{1}{v^+ + v^-}\sum_{ij}\left(w_{ij}^- - e_{ij}^-\right)\delta_{M_iM_j}$$

The community assignment for each region was then estimated 500 times across a range of γ values (0.5–2.0, in steps of 0.1). In order to identify multi-level structure in our data, we repeated the modularity analysis for each of the modules identified in the first step[104]. Finally, a consensus partition was identified using a fine-tuning algorithm from the Brain Connectivity Toolbox (*http://www.brain-connectivity-toolbox.net/*). We subsequently used this final module assignment to estimate the cartographic profile of the each participant's time-averaged adjacency matrix[80]. Specifically, we estimated Integration using the participation coefficient, which quantifies the extent to which a region connects across all modules (i.e., between-module strength[105]), and Segregation using the module-degree Z-score. These measures were entered into a joint-histogram (101×101 unique bins, equally-spaced between 0–1 [for Integration] and - 1 and 1 [for Segregation]). The value within each bin of this joint histogram was then correlated with the combined regression weights of PC_2 and PC_3 for each subject. A permutation test that scrambled the order of participants was used to assess statistical significance ($p < 0.05$).

Brain-state displacement and the energy landscape

To quantify the change in the evoked BOLD activity following each stimulus we calculated the main BOLD displacement (MBD). The MBD is a measure of the absolute evoked deviation in BOLD activity. The evoked activity is measured through a general linear model using a canonical hemodynamic response function convolved on a design matrix. We are interested in the probability, p(MBD, *re*), that we will observe a given displacement in BOLD at a given regressor *re*. The probability is calculated through the null model of the general lineal model (the probability that the observed evoked value of the corresponding region is different from 0). As described above we then calculated the energy for each displacement value as $E_{MBD,\,re} = \mathrm{In}\left(\frac{1}{P(MBD,re)}\right)$.

Finally, to measure the surprise per displacement, we divided the absolute β for PC2 from $E_{MBD,\,re}$ for each regressor *re* (**Fig. 6**).

Supplementary Files

Figure S1.

Overall Analysis Flow.

Top Row (orange) - pupil diameter was collected in a cohort of 35 individuals while they performed the Ambiguous Figures task. We observed a large peak in pupil dilation at the perceptual change point, which led us to make the prediction that there should be an increase in inter-regional gain at the switch point. Middle Row (blue) - we trained a 100-node RNN to perform a similar classification task in the presence of shifting perceptual ambiguity, and then tested the network at different levels of gain (i.e., the slope of the *tanh* activation function). We observed early switches with heightened gain, as well as altered attractor dynamics that caused a flattening of the energy landscape characterising state switches. Bottom Row (green) - we tested the predictions of the RNN using BOLD data from 17 subjects performing the same task. After filtering the BOLD data through a principal component analysis (in which we retained the top 5 principal components; PC_{1-5}), we observed an increase in the gradient of PC loading around the switch point using an FIR model, as well as a flattening of the energy landscape, thus confirming our original predictions.

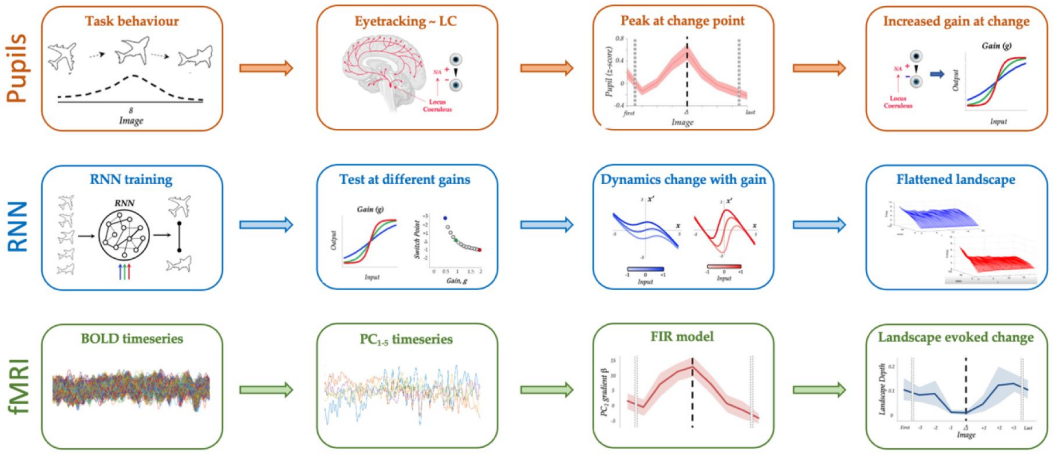
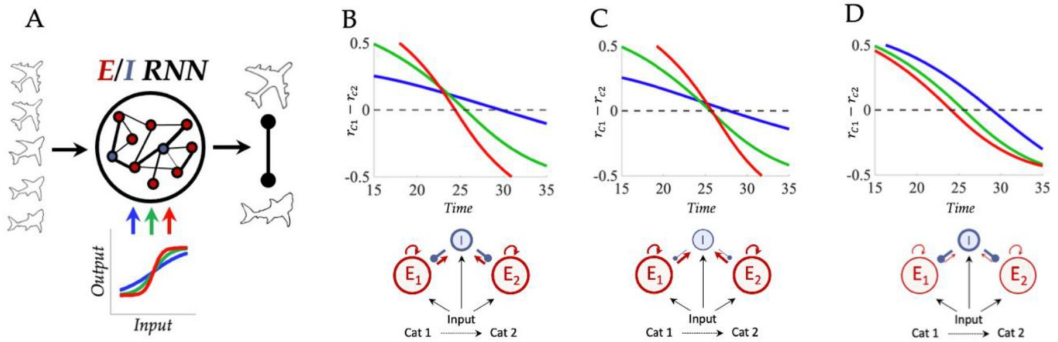


Figure S2.

Difference in mean firing rates between stimulus selective excitatory clusters.

To examine the effect of manipulating gain on the operation of the network (**A**) we averaged over the firing rate of the excitatory neurons in each stimulus selective cluster (r_{c1}, r_{c2}) and looked for the point at which $r_{c2} > r_{c1}$ (and v.v.). In line with expectations the speeding and slowing effect of gain on network output time was straightforwardly reflected in the mean firing rates. **B)** Difference in mean firing rates for high (red), intermediate (green), and low (blue) gain, for gain manipulations targeting both excitatory and inhibitory neurons. Notice that the switch from $r_{c1} > r_{c2}$ to $r_{c2} > r_{c1}$ occurs sooner in time for high gain, and slower for low gain. **C)** Manipulating excitatory gain in isolation led to slower switches in mean firing rates for low gain but high gain did not speed switches. **D)** Manipulating inhibitory gain in isolation led to slower switches in mean firing rates for low gain and speeded switches under high gain.



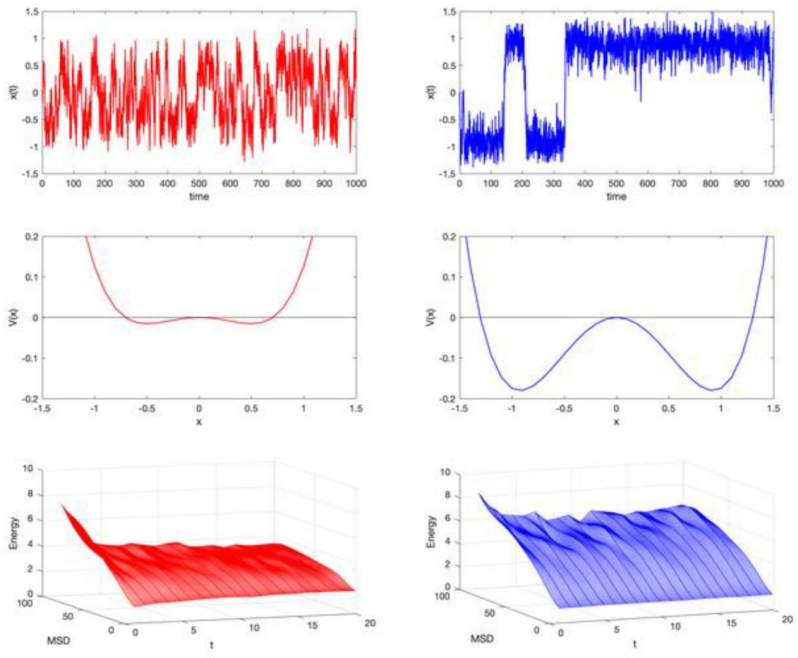


Figure S3.

Relationship between Attractor Depth and Energy Landscape.

We simulated a simple model $\frac{dx}{dt} = \alpha x - x^3$ (the normal form of a pitchfork bifurcation) (middle row is the corresponding potential $\frac{dx}{dt} = -\frac{dv}{dx}$) and set the α term to two different values: on the left (red), α = 0.25, which corresponds to relatively shallow attractors; on the right (blue), α = 0.75, which corresponds to relatively deep attractors. We simulated model in the presence of light noise noisy, and then calculated the energy landscape of the timeseries (see methods in main paper) (bottom row; *z*-axis) As can be seen by comparing the middle and bottom rows, deeper attractors relate to higher energy barriers.

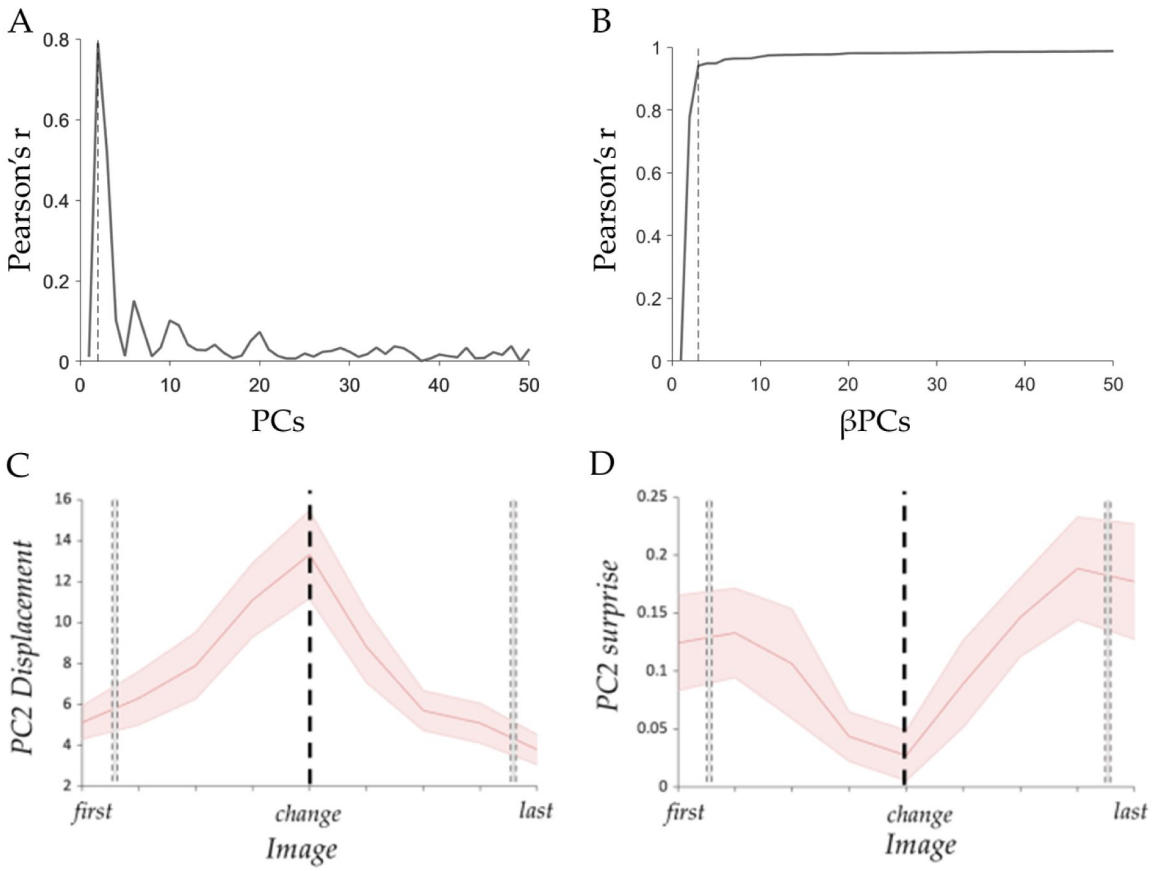


Figure S4

PC2 evoked displacement and surplice around the perceptual change.

A) Pearson's correlation between each PCs and the evoked brain activity at the perceptual switch (β values), dashed line at PC2. B) Pearson's correlation between the inverted brain maps using βPC ($PC_{(i-i)} \times \beta PC_{(i-i)}$). Dashed line shows that the correlation gets to 94% using the first 3 PCs (Pearson's r =0.94, $p < 0.001$). C) Mean absolute β loading (red) and group standard error (shaded red). D) Mean surprise calculated as -log(1-pvalues) in each regressors. Dotted black line define the perceptual switch point.

References

1. Bogacz R. 2017) **A tutorial on the free-energy framework for modelling perception and learning** *J. Math. Psychol* **76**:198–211

2. Flounders M. W., González-Garcia C., Hardstone R., He B. J. 2019) **Neural dynamics of visual ambiguity resolution by perceptual prior** *eLife* **8**:e41861

3. Friston K. 2005) **A theory of cortical responses** *Philos. Trans. R. Soc. B Biol. Sci* **360**:815–836

4. Hohwy J. 2013) **The Predictive Mind** OUP Oxford

5. Clark A. 2013) **Whatever next? Predictive brains, situated agents, and the future of cognitive science** *Behav. Brain Sci* **36**:181–204

6. Hohwy J., Roepstorff A., Friston K. 2008) **Predictive coding explains binocular rivalry: An epistemological review** *Cognition* **108**:687–701

7. Alais D., Blake R. 2005) **Binocular Rivalry** MIT Press

8. van Ee R. 2005) **Dynamics of perceptual bi-stability for stereoscopic slant rivalry and a comparison with grating, house-face, and Necker cube rivalry** *Vision Res* **45**:29–40

9. Hohwy J. 2012) **Attention and Conscious Perception in the Hypothesis Testing Brain** *Front. Psychol* **3**

10. Moran R. J., et al. 2013) **Free Energy, Precision and Learning: The Role of Cholinergic Neuromodulation** *J. Neurosci* **33**:8227–8236

11. Shine J. M., et al. 2021) **Computational models link cellular mechanisms of neuromodulation to large-scale neural dynamics** *Nat. Neurosci* **24**:765–776

12. Parr T., Friston K. J. 2018) **The Anatomy of Inference: Generative Models and Brain Structure** *Front. Comput. Neurosci* **12**:90

13. Parr T., Friston K. J. 2017) **Uncertainty, epistemics and active inference** *J. R. Soc. Interface* **14**:20170376

14. Thura D., Cabana J.-F., Feghaly A., Cisek P. 2020) **Unified Neural Dynamics of Decisions and Actions in the Cerebral Cortex and Basal Ganglia** https://doi.org/10.1101/2020.10.22.350280

15. Bogacz R., Brown E., Moehlis J., Holmes P., Cohen J. D. 2006) **The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks** *Psychol. Rev* **113**:700–765

16. Murphy P. R., Boonstra E., Nieuwenhuis S. 2016) **Global gain modulation generates timedependent urgency during perceptual choice in humans** *Nat. Commun* **7**:1–15

17. Stöttinger E., et al. 2015) **A cortical network that marks the moment when conscious representations are updated** *Neuropsychologia* **79**:113–122

18. Weilnhammer V., Stuke H., Hesselmann G., Sterzer P., Schmack K. 2017) **A predictive coding account of bistable perception - a model-based fMRI study** *PLoS Comput. Biol* **13**

19. Reynolds J. H., Heeger D. J. 2009) **The Normalization Model of Attention** *Neuron* **61**:168–185

20. Desimone R., Duncan J. 1995) **Neural mechanisms of selective visual attention** *Annu. Rev. Neurosci* **18**:193–222

21. Meng M., Tong F. 2004) **Can attention selectively bias bistable perception? Differences between binocular rivalry and ambiguous figures** *J. Vis* **4**:539–551

22. Dieter K. C., Tadin D. 2011) **Understanding attentional modulation of binocular rivalry: A framework based on biased competition** *Front. Hum. Neurosci* **5**:1–12

23. Wong K.-F. 2006) **A Recurrent Network Mechanism of Time Integration in Perceptual Decisions** *J. Neurosci* **26**:1314–1328

24. Eckhoff P., Wong-Lin K., Holmes P. 2011) **Dimension Reduction and Dynamics of a Spiking Neural Network Model for Decision Making under Neuromodulation** *SIAM J. Appl. Dyn. Syst* **10**:148–188

25. Wang X.-J. 2002) **Probabilistic Decision Making by Slow Reverberation in Cortical Circuits** *Neuron* **36**:955–968

26. Cisek P. 2019) **Resynthesizing behavior through phylogenetic refinement** *Atten. Percept. Psychophys* **81**:2265–2287 https://doi.org/10.3758/s13414-019-01760-1

27. Carter O., van Swinderen B., Leopold D. A., Collin S. P., Maier A. 2020) **Perceptual rivalry across animal species** *J. Comp. Neurol* **528**:3123–3133

28. Sales A. C., Friston K. J., Jones M. W., Pickering A. E., Moran R. J. 2019) **Locus Coeruleus tracking of prediction errors optimises cognitive flexibility: An Active Inference model** *PLOS Comput. Biol* **15**:e1006267

29. Vincent P., Parr T., Benrimoh D., Friston K. J. 2019) **With an eye on uncertainty: Modelling pupillary responses to environmental volatility** *PLOS Comput. Biol* **15**:e1007126

30. Jordan R., Keller G. B. 2023) **The locus coeruleus broadcasts prediction errors across the cortex to promote sensorimotor plasticity** *eLife* **12**

31. Murphy P. R., Vandekerckhove J., Nieuwenhuis S. 2014) **Pupil-Linked Arousal Determines Variability in Perceptual Decision Making** *PLoS Comput. Biol* **10**

32. Szabadi E. 2018) **Functional Organization of the Sympathetic Pathways Controlling the Pupil: Light-Inhibited and Light-Stimulated Pathways** *Front. Neurol* **9**

33. Briand L. A., Gritton H., Howe W. M., Young D. A., Sarter M. 2007) **Modulators in concert for cognition : Modulator interactions in the prefrontal cortex** :69–91

34. Sara S. J. 2009) **The locus coeruleus and noradrenergic modulation of cognition** *Nat. Rev. Neurosci* **10**:211–223

35. Jacob S. N., Nienborg H., Sara S. J., Jacob S. N. 2018) **Monoaminergic Neuromodulation of Sensory Processing** *Front Neural Circuits.* :1–17

36. Bouret S., Sara S. J. 2005) **Network reset: A simplified overarching theory of locus coeruleus noradrenaline function** *Trends Neurosci* **28**:574–582

37. Joshi S., Gold J. I. 2020) **Pupil Size as a Window on Neural Substrates of Cognition** *Trends Cogn. Sci* **24**:466–480

38. Samuels E., Szabadi E. 2008) **Functional Neuroanatomy of the Noradrenergic Locus Coeruleus: Its Roles in the Regulation of Arousal and Autonomic Function Part II: Physiological and Pharmacological Manipulations and Pathological Alterations of Locus Coeruleus Activity in Humans** *Curr. Neuropharmacol* **6**:254–285

39. Pfeffer T., et al. 2022) **Coupling of pupil- and neuronal population dynamics reveals diverse influences of arousal on cortical processing** *eLife* **11**:e71890

40. de Gee J. W., et al. 2020) **Pupil-linked phasic arousal predicts a reduction of choice bias across species and decision domains** *eLife* **9**:e54014

41. Einhäuser W., Stout J., Koch C., Carter O. 2008) **Pupil dilation reflects perceptual selection and predicts subsequent stability in perceptual rivalry** *Proc. Natl. Acad. Sci. U. S. A* **105**:1704–1709

42. Reimer J., et al. 2016) **Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex** *Nat. Commun* **7**:1–7

43. Shine J. M. 2019) **Neuromodulatory Influences on Integration and Segregation in the Brain** *Trends Cogn. Sci* **23**:572–583

44. Wainstein G., et al. 2021) **The ascending arousal system promotes optimal performance through meso-scale network integration in a visuospatial attentional task** *Netw. Neurosci* **5**:890–910 https://doi.org/10.1162/netn_a_00205

45. de Gee J. W., Knapen T., Donner T. H. 2014) **Decision-related pupil dilation reflects upcoming choice and individual bias** *Proc. Natl. Acad. Sci* **111**:E618-E625

46. de Gee J. W., et al. 2017) **Dynamic modulation of decision biases by brainstem arousal systems** *eLife* **6**:e23232

47. Munn B. R., Müller E. J., Wainstein G., Shine J. M. 2021) **The ascending arousal system shapes neural dynamics to mediate awareness of cognitive states** *Nat. Commun* **12**:1–9

48. Stöttinger E., Sepahvand N. M., Danckert J., Anderson B. 2016) **Assessing perceptual change with an ambiguous figures task: Normative data for 40 standard picture sets** *Behav. Res. Methods* **48**:201–222

49. Shine J. M., Aburn M. J., Breakspear M., Poldrack R. A. 2018) **The modulation of neural gain facilitates a transition between functional segregation and integration in the brain** *eLife* **7**:e31130

50. Servan-schreiber A. D., et al. 2016) **Reports A Network Model of Catecholamine Effects : Gain , Signal-to-Noise Ratio , and Behavior** :892–895

51. Joshi S., Li Y., Kalwani R. M., Gold J. I. 2016) **Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex** *Neuron* **89**:221–234

52. Hupe J. M., Lamirel C., Lorenceau J. 2009) **Pupil dynamics during bistable motion perception** *J. Vis* **9**:10–10

53. Kloosterman N. A., et al. 2015) **Pupil size tracks perceptual content and surprise** *Eur. J. Neurosci* **41**:1068–1078

54. Kosciessa J. Q., Lindenberger U., Garrett D. D. 2021) **Thalamocortical excitability modulation guides human perception under uncertainty** *Nat. Commun* **12**

55. Eldar E., Cohen J. D., Niv Y. 2013) **The effects of neural gain on attention and learning** *Nat. Neurosci* **16**:1146–1153

56. Urai A. E., Braun A., Donner T. H. 2017) **Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias** *Nat. Commun* **8**:14637

57. Kloosterman N. A., et al. 2015) **Pupil size tracks perceptual content and surprise** *Eur. J. Neurosci* **41**:1068–1078

58. Müller E. J., Munn B. R., Shine J. M. 2020) **Diffuse neural coupling mediates complex network dynamics through the formation of quasi-critical brain states** *Nat. Commun* **11**

59. Shine J. M., et al. 2016) **The Dynamics of Functional Brain Networks: Integrated Network States during Cognitive Task Performance** *Neuron* **92**:544–554

60. Taylor N. L., et al. 2022) **Structural connections between the noradrenergic and cholinergic system shape the dynamics of functional brain networks** *NeuroImage* **260**:119455

61. Aston-Jones G., Cohen J. D. 2005) **An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance** *Annu. Rev. Neurosci* **28**:403–450

62. Song H. F., Yang G. R., Wang X.-J. 2016) **Training Excitatory-Inhibitory Recurrent Neural Networks for Cognitive Tasks: A Simple and Flexible Framework** *PLOS Comput. Biol* **12**:e1004792

63. Yang G. R., Wang X. J. 2020) **Artificial Neural Networks for Neuroscientists: A Primer** *Neuron* **107**:1048–1070

64. Song H. F., Yang G. R., Wang X.-J. 2016) **Training Excitatory-Inhibitory Recurrent Neural Networks for Cognitive Tasks: A Simple and Flexible Framework** *PLOS Comput. Biol* **12**:e1004792

65. Wainstein G., Müller E. J., Taylor N., Munn B., Shine J. M. 2022) **The role of the locus coeruleus in shaping adaptive cortical melodies** *Trends Cogn. Sci* **26**:527–538

66. Barack D. L., Krakauer J. W. 2021) **Two views on the cognitive brain** *Nat. Rev. Neurosci* **22**:359–371

67. Beer R. D. 2022) **Codimension-2 parameter space structure of continuous-time recurrent neural networks** *Biol. Cybern* **116**:501–515

68. Beer R. D. 2000) **Dynamical approaches to cognitive science** *Trends Cogn. Sci* **4**:91–99

69. Sussillo D. 2014) **Neural circuits as computational dynamical systems** *Curr. Opin. Neurobiol* **25**:156–163

70. Sussillo D., Barak O. 2013) **Opening the Black Box: Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks** *Neural Comput* **25**:626–649

71. Taylor N. L., et al. 2022) **NeuroImage Structural connections between the noradrenergic and cholinergic system shape the dynamics of functional brain networks**

72. John Y. J., et al. 2022) **It's about time: Linking dynamical systems with human neuroimaging to understand the brain** *Netw. Neurosci* **6**:960–979

73. Richards B. A., et al. 2019) **A deep learning framework for neuroscience** *Nat. Neurosci* **22**:1761–1770

74. Doerig A., et al. 2023) **The neuroconnectionist research programme** *Nat. Rev. Neurosci* **24**:431–450 https://doi.org/10.1038/s41583-023-00705-w

75. Vyas S., Golub M. D., Sussillo D., Shenoy K. V. 2020) **Computation Through Neural Population Dynamics** *Annu. Rev. Neurosci* **43**:249–275

76. Shine J. M., et al. 2019) **The low-dimensional neural architecture of cognitive complexity is related to activity in medial thalamic nuclei** *Neuron* **104**:849–855

77. Jolliffe I. T. 1982) **A Note on the Use of Principal Components in Regression** *Appl. Stat* **31**:300

78. Yarkoni T., Poldrack R. A., Nichols T. E., Van Essen D. C., Wager T. D. 2011) **Large-scale automated synthesis of human functional neuroimaging data** *Nat. Methods* **8**:665–670

79. Shine J. M., Aburn M. J., Breakspear M., Poldrack R. A. 2018) **The modulation of neural gain facilitates a transition between functional segregation and integration in the brain** *eLife* **7**:e31130

80. Shine J. M., et al. 2016) **The Dynamics of Functional Brain Networks: Integrated Network States during Cognitive Task Performance** *Neuron* **92**:544–554

81. Sara S. J. 2015) **Locus Coeruleus in time with the making of memories** *Curr. Opin. Neurobiol* **35**:87–94

82. Liu Y., Rodenkirch C., Moskowitz N., Schriver B. 2017) **Dynamic Lateralization of Pupil Dilation Evoked by Locus Coeruleus Activation Results from Article Dynamic Lateralization of Pupil Dilation Evoked by Locus Coeruleus Activation Results from Sympathetic , Not Parasympathetic , Contributions** *CellReports* **20**:3099–3112

83. Nieuwenhuis S., Aston-Jones G., Cohen J. D. 2005) **Decision making, the P3, and the locus coeruleus-norepinephrine system** *Psychol. Bull* **131**:510–532

84. Totah N. K. B., Logothetis N. K., Eschenko O. 2018) **Noradrenergic ensemble-based modulation of cognition over multiple timescales** *Brain Res* **1709**:50–66 https://doi.org/10.1016/j.brainres.2018.12.031

85. Zerbi V., et al. 2019) **Rapid Reconfiguration of the Functional Connectome after Chemogenetic Locus Coeruleus Activation** *SSRN Electron. J* https://doi.org/10.2139/ssrn.3334983

86. Hansen J. Y., et al. 2022) **Mapping neurotransmitter systems to the structural and functional organization of the human neocortex** *Nat. Neurosci* **25**:1569–1581

87. Hansen J. Y., et al. 2021) **Mapping gene transcription and neurocognition across human neocortex** *Nat. Hum. Behav* **5**:1240–1250

88. Cazettes F., Reato D., Morais J. P., Renart A., Mainen Z. F. 2020) **Phasic Activation of Dorsal Raphe Serotonergic Neurons Increases Pupil Size** *Curr. Biol* **31**:192–197 https://doi.org/10.1016/j.cub.2020.09.090

89. Alnæs D., et al. 2014) **Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus** *J. Vis* **14**:1–20

90. Janitzky K., Lippert M. T., Engelhorn A., Tegtmeier J., Cope Z. A. 2015) **Optogenetic silencing of locus coeruleus activity in mice impairs cognitive flexibility in an attentional set-shifting task** :1–8

91. Reimer J., et al. 2014) **Pupil Fluctuations Track Fast Switching of Cortical States during Quiet Wakefulness** *Neuron* **84**:355–362

92. Wainstein G., et al. 2017) **Pupil size tracks attentional performance in attention-deficit/hyperactivity disorder** *Sci. Rep* **7**:1–9

93. Campos-Arteaga G., et al. 2020) **Differential neurophysiological correlates of retrieval of consolidated and reconsolidated memories in humans: an ERP and pupillometry study** *Neurobiol. Learn. Mem* :107279 https://doi.org/10.1016/j.nlm.2020.107279

94. Rojas-Líbano D., et al. 2019) **A pupil size, eye-tracking and neuropsychological dataset from ADHD children during a cognitive task** *Sci. Data* **6**:25

95. Paszke A., et al. 2019) **Pytorch: An imperative style, high-performance deep learning library** *Adv. Neural Inf. Process. Syst* **32**

96. Werbos P. J. 1990) **Backpropagation Through Time: What It Does and How to Do It** *Proc. IEEE* **78**:1550–1560

97. Kingma D. P., Ba J. L. 2015) **Adam: A method for stochastic optimization** *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc* :1–15

98. Tkacik G., et al. 2015) **Thermodynamics and signatures of criticality in a network of neurons** *Proc. Natl. Acad. Sci. U. S. A* **112**:11508–11513

99. Power J. D., et al. 2014) **Methods to detect, characterize, and remove motion artifact in resting state fMRI** *NeuroImage* **84**:320–41

100. Behzadi Y., Restom K., Liau J., Liu T. T. 2007) **A component based noise correction method (CompCor) for BOLD and perfusion based fMRI** *NeuroImage* **37**:90–101

101. Gu S., et al. 2015) **Controllability of structural brain networks** *Nat. Commun* **6**:1–10

102. Gordon E. M., et al. 2016) **Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations** *Cereb. Cortex* **26**:288–303

103. Diedrichsen J., Balsters J. H., Flavell J., Cussans E., Ramnani N. 2009) **A probabilistic MR atlas of the human cerebellum** *NeuroImage* **46**:39–46

104. Meunier D., Lambiotte R., Bullmore E. T. 2010) **Modular and Hierarchically Modular Organization of Brain Networks** *Front. Neurosci* **4**

105. Guimerà R., Nunes Amaral L. A. 2005) **Functional cartography of complex metabolic networks** *Nature* **433**:895–900

Author information

Gabriel Wainstein[†]
Brain and Mind Center, The University of Sydney, Sydney, Australia
ORCID iD: 0000-0002-8106-6647

[†]Co-first author

Christopher J Whyte[†]
Brain and Mind Center, The University of Sydney, Sydney, Australia, Center for Complex Systems, The University of Sydney, Sydney, Australia

[†]Co-first author

Kaylena A Ehgoetz Martens
The University of Waterloo, Waterloo, Canada

Eli J Müller
Brain and Mind Center, The University of Sydney, Sydney, Australia, Center for Complex Systems, The University of Sydney, Sydney, Australia

Vicente Medel
Brain and Mind Center, The University of Sydney, Sydney, Australia, Latin American Brain Health (BrainLat), Universidad Adolfo Ibanez, Santiago, Chile

Britt Anderson
The University of Waterloo, Waterloo, Canada

Elisabeth Stöttinger
Hochschule Fresenius, Idstein, Germany

James Danckert
The University of Waterloo, Waterloo, Canada

Brandon R Munn[ζ]
Brain and Mind Center, The University of Sydney, Sydney, Australia, Center for Complex Systems, The University of Sydney, Sydney, Australia
ORCID iD: 0000-0002-3638-1605

[ζ]Co-senior author

James M Shine[ζ]
Brain and Mind Center, The University of Sydney, Sydney, Australia, Center for Complex Systems, The University of Sydney, Sydney, Australia
ORCID iD: 0000-0003-1762-5499

For correspondence: mac.shine@sydney.edu.au
[ζ]Co-senior author

Editors

Reviewer #1 (Public review):

Summary:

This paper proposes a neural mechanism underlying the perception of ambiguous images: neuromodulation changes the gain of neural circuits promoting a switch between two possible percepts. Converging evidence for this is provided by indirect measurements of neuromodulatory activity and large-scale brain dynamics which are linked by a neural network model. However, both the data analysis as well as the computational modeling are incomplete and would benefit from a more rigorous approach.

This is a revised version of the manuscript which, in my view, is a considerable step forward compared to the original submission.

In particular, the authors now model phasic gain changes in the RNN, based on the network's uncertainty. This is original and much closer to what is suggested by the phasic pupil responses. They also show that switching is actually a network effect because switching times depend on network configuration (Fig 2). This resolves my main comments 1 and 2 about the model.

The mechanism, as I understand it, is different from what the authors described before in the RNN with tonic gain changes. As uncertainty increases, the network enters a regime in which the two excitatory populations start to oscillate. My intuition is that this oscillation arises from the feedback loop created by the new gain control mechanism. If my intuition is correct, I think it would be worth to explain this mechanism in the paper more explicitly.

Overall, the modeling part of the paper has changed quite a lot and I think it is now more solid which is why I have updated my "strength of evidence" rating.

https://doi.org/10.7554/eLife.93191.2.sa2

Reviewer #2 (Public review):

This paper tests the hypothesis that perceptual switches during the presentation of ambiguous stimuli are accompanied by changes in neuromodulation that alter neural gain and trigger abrupt changes in brain activity. To test this hypothesis, the study combines pupillometry, artificial recurrent network (RNN) analysis and fMRI recording. In particular, the study uses methods of energy landscape analysis inspired by physics, which is particularly interesting.

Strengths

- The authors should be commended for combining different methods (pupillometry, RNNs, fMRI) to test their hypothesis. This combination provides a mechanistic insight into perceptual switches in the brain and artificial neural networks.
- The study combines different viewpoints and fields of scientific literature, including neuroscience, psychology, physics, and dynamical systems. In order to make this combination more accessible to the reader, the different aspects are presented in a pedagogical way to be accessible to all fields.
- This combination of methods and viewpoints is rarely done, so it is very useful.
- The authors introduce dynamic gain modulation in their recurrent neural network, which is novel. They devote a section of the paper to studying the dynamics, fixed points and convergence of this type of network.

Weaknesses

- The study may not be specific to perceptual switches. This is because the study relies on a paradigm in which participants report when they identify a switch in the item category. Therefore, it is unclear whether the effects reported in the paper are related to the perceptual switch itself, to attention, or to the detection of behaviourally relevant events. The authors are cautious and explicitly acknowledge this point in their study.
- The demonstration of the causal role of gain modulation in perceptual switches is partial. This causality is clearly demonstrated in the simulation work with the RNN. However, it is not fully demonstrated in the pupil analysis and the fMRI analysis. One reason is that this work is correlative (which is already very informative). An analysis of the timing of the effect might have overcome this limitation. For example, in a previous study, the same group showed that fMRI activity in the LC region precedes changes in the energy landscape of fMRI dynamics, which is a step towards investigating causal links between gain modulation, changes in the energy landscape and perceptual switches.
- Some effects may reflect the expectation of a perceptual switch rather than the perceptual switch itself. To mitigate this risk, the design of the fMRI task included catch trials, in which no switch occurs, to reduce the expectation of a switch. The pupil study, however, did not include such catch trials.
- The paper uses RNN-based modelling to provide mechanistic insight into the role of gain modulation in perceptual switches. However, the RNN solves a task that differs markedly from that performed by human participants, which may limit the explanatory value of the model. The RNN is provided with two inputs characterising the sensory evidence supporting the first and last image category in the sequence (e.g. plane and shark). In contrast, observers in the task were naïve as to the identity of the last image at the beginning of the sequence. The brain first receives sensory evidence about the image category (e.g. plane) with which the sequence begins, which is very easy to recognise, then it sees a sequence of morphed images and has to discover what the final image category will be. To discover the final image category, the brain has to search a vast space of possible second images (it is a shark?, a frog?, a bird?, etc.), rather than comparing the likelihood of just two categories. This search process

and the perceptual switch in the task appear to be mechanistically different from the competition between two inputs in the RNN.
- Another aspect of the motivation for the RNN model remains unclear. The authors introduce dynamic gain modulation in the RNN, but it is not clear what the added value of dynamic gain modulation is. Both static (Fig. S1) and dynamic (Fig. 2F) gain modulation lead to the predicted effect: faster switching when the gain is larger.
- The authors are to be commended for addressing their research questions with multiple tools and approaches. There are links between the different parts of the study. The RNN and the pupil are linked by the notion of gain modulation, the RNN and the fMRI analysis are linked by the study of the energy landscape, the fMRI study and the pupil study are indirectly linked by previous work for this group showing that the peak in LC fMRI activity precedes a flattening of the energy landscape. These links are very interesting but could have been stronger and more complete.

https://doi.org/10.7554/eLife.93191.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *Summary:*
>
> *This paper investigates the neural mechanisms underlying the change in perception when viewing ambiguous figures. Each possible percept is related to an attractor-like brain state and a perceptual switch corresponds to a transition between these states. The hypothesis is that these switches are promoted by bursts of noradrenaline that change the gain of neural circuits. The authors present several lines of evidence consistent with this view: pupil diameter changes during the time point of the perceptual change; a gain change in neural network models promotes a state transition; and large-scale fMRI dynamics in a different experiment suggests a lower barrier between brain states at the change point. However, some assumptions of the computational model seem not well justified and the theoretical analysis is incomplete. The paper would also benefit from a more in-depth analysis of the experimental data.*
>
> *Strengths:*
>
> *The main strength of the paper is that it attempts to combine experimental measurements - from psychophysics, pupil measurements, and fMRI dynamics - and computational modeling to provide an emerging picture of how a perceptual switch emerges. This integrative approach is highly useful because the model has the potential to make the underlying mechanisms explicit and to make concrete predictions.*
>
> *Weaknesses:*
>
> *A general weakness is that the link between the three parts of the paper is not very strong. Pupil and fMRI measurements come from different experiments and additional analysis showing that the two experiments are comparable should be included. Crucially, the assumptions underlying the RNN modeling are unclear and the conclusions drawn from the simulation may depend on those assumptions.*

With this comment in mind we have made substantial effort to better integrate the three different aspects of our paper. On the pupillometry side, we now show that the dynamic

uncertainty associated with perceptual categorisation shares a similar waveform with the observed fluctuations in pupil diameter around the switch point (Fig 2B). To better link the modelling to the behaviour we have also made the gain of the activation function of each sigmoidal unit change dynamically as a function of the uncertainty (i.e. the entropy) of the network's classification generating phasic changes in gain that mimic the observed phasic changes in pupil dilation explicitly linking the dynamics of gain in the RNN to the observed dynamics of pupil diameter (our non-invasive proxy for neuromodulatory tone). Finally we note that the predictions of the RNN (flattened egocentric landscape and peaks in low-dimensional brain state velocity at the time point of the perceptual switch) were tested directly in the whole-brain BOLD data, which links the modelling and BOLD analysis. Finally we note that whilst we agree that an experiment in which pupilometry and BOLD data were collected simultaneously would be ideal, these data were not available to us at the time of this study.

> *Main points:*
>
> *Perceptual tasks in pupil and fMRI experiments: how comparable are these two tasks? It seems that the timing is very different, with long stimulus presentations and breaks in the fMRI task and a rapid sequence in the pupil task. Detailed information about the task timing in the pupil task is missing. What evidence is there that the same mechanisms underlie perceptual switches at these different timescales? Quantification of the distributions of switching times/switching points in both tasks is missing. Do the subjects in the fMRI task show the same overall behavior as in the pupil task? More information is needed to clarify these points.*

We recognize the need for a more detailed and comparative analysis of the perceptual tasks used in our pupil and fMRI experiments, particularly regarding differences in timing, task structure, and instructions. The fMRI task incorporates jittered inter-trial intervals (ITIs) of 2, 4, 6, and 8 seconds, designed to enable effective deconvolution of the BOLD response (Stottinger et al., 2018). In contrast, the pupil task presents a more rapid sequence of stimuli without ITIs. These timing differences are reflected in the mean perceptual switch points: the 8th image in the fMRI task and the 9th image in the pupil task. This small yet consistent difference suggests subtle influences of task design on behavior.

Despite these structural and instructional differences, our analyses indicate that overall behavioral patterns remain consistent across the two modalities. The distributions of switching times align closely, and no significant behavioral deviations were observed that might suggest a fundamental difference in the underlying mechanisms driving perceptual switches. These findings suggest that the additional time and structural differences in the fMRI task do not significantly alter the behavioral outcomes compared to the pupil task.

To address these issues, we have added paragraphs in the Results, Methods, and Limitations sections of the manuscript. In the Results section, we provide a detailed comparison of switching point distributions across the two tasks, emphasizing behavioral consistencies and any observed variations. In the Methods section, we include an expanded description of task timing, instructions, and the presence or absence of catch trials to ensure clarity regarding the experimental setups. Finally, in the Limitations section, we acknowledge the structural differences between the tasks, particularly the lack of catch trials and rapid stimulus presentation in the pupil task, and discuss how these differences may influence perceptual dynamics.

These additions aim to clarify how task-specific factors, such as timing, instructions, and catch trials, influence perceptual dynamics while highlighting the consistency in behavioral outcomes across both experimental setups. We believe these revisions address the concerns raised and enhance the manuscript's transparency and rigor.

> *Computational model:*
>
> *(1) Modeling noradrenaline effects in the RNN: The pupil data suggests phasic bursts of NA would promote perceptual switches. But as I understand, in the RNN neuromodulation is modeled as different levels of gain throughout the trial. Making the neural gain time-dependent would allow investigation of whether a phasic gain change can explain the experimentally observed distribution of switching times.*

We thank the reviewer for this very helpful suggestion. We updated the RNN so that, post-training, gain changes dynamically as a function of the network's classification uncertainty (i.e. the entropy of the network's output). Specifically, the gain dynamics of each unit in the neural network are governed by a linear ODE with a forcing function given by the entropy of the network's classification (i.e. the uncertainty of the classification). This explicitly tests the hypothesis that uncertainty driven increases in gain near the perceptual switch (when the input is maximally ambiguous) speeds perceptual switches, and allows us to distinguish between tonic and phasic increases in gain (in the absence of uncertainty forcing gain decays exponentially to a tonic value of 1). Importantly, in line with our hypothesis, we found that switch times decreased as we increased the impact of uncertainty on gain (i.e. switch times decreased as the magnitude of uncertainty forcing increased). Finally, we wish to note that although making gain dynamical is relatively simple conceptually, actually implementing it and then analysing the dynamics turned out to be highly non-trivial. To our knowledge our model is the first RNN of reasonable size to implement dynamical gain requiring us to push the RNN modelling beyond the current state of the art (see Fig 2 - 4).

> *(2) Modeling perceptual switches: in the results, it is described that the networks were trained to output a categorical response, but the firing rates in Fig 2B do not seem categorical but rather seem to follow the input stimulus. The output signals of the network are not shown. If I understand correctly, a trivial network that would just represent the two input signals without any internal computation and relay them to the output would do the task correctly (because "the network's choice at each time point was the maximum of the two-dimensional output", p. 22). This seems like cheating: the very operation that the model should perform is to signal the change, in a categorical manner, not to represent the gradually changing input signals.*

The output of the network was indeed trained to be categorical via a cross entropy loss function with the output defined by the max of the projection of the excitatory hidden units onto the output weights which is boilerplate RNN modelling practice. As requested we now show the output in Fig 2B. On the broader question of whether a trivially small network could solve the task we are in total agreement that with the right set of hand-crafted weights a two neuron sigmoidal network with winner-take-all readout could solve the task. We disagree, however, that using an RNN is cheating in any way. Many tasks in neuroscience can be trivially solved with a very small number of recurrent units (e.g. basically all 2AF tasks). The question we were interested in is how the brain might solve the task, and more specifically how neuromodulator control of gain changes the dynamics of our admittedly very simple task. We could have done this by hand crafting a small network to solve the task but we wanted to use the RNN modelling as a means of both hypothesis testing and hypothesis generation. We now expand on and justify this modelling choice in the second paragraph of the discussion:

"We chose to use an RNN, instead of a simpler (more transparent) model as we wanted to use the RNN as a means of both hypothesis generation and hypothesis testing. Specifically, unlike more standard neuronal models which are handcrafted to reproduce a specific effect, when building an RNN the modeller only specifies the network inputs, labels, and the parameter constraints (e.g. Dale's law) in advance. The dynamics of the RNN are entirely determined by

optimisation. Post-training manipulations of the RNN are not built in, or in any way guaranteed to work, making them more analogous to experimental manipulations of an approximately task-optimal brain-like system. Confirmatory results are arguably, therefore, a first steps towards an in vitro experimental test."

> *(3) The mechanism of how increased gain leads to faster switches remains unclear to me. My first intuition was that increasing the gain of excitatory populations (the situation shown in Fig. 2E) in discrete attractor models would lead to deeper attractor wells and this would make it more difficult to switch. That is, a higher gain should lead to slower decisions in this case. However, here the switching time remains constant for a gain between 1 and 1.5. Lowering the gain, on the other hand, leads to slower switching. It is, of course, possible that the RNN behaves differently than classical point attractor models or that my intuition is incorrect (though I believe it is consistent with previous literature, e.g. Niyogi & Wong-Lin 2013 (doi:10.1371/journal.pcbi.1003099) who show higher firing rates - more stable attractors - for increased excitatory gain).*

We thank the reviewer for the astute observation, which we entirely agree with. The energy landscape analysis is a method still under active development within our group and we are still learning how to best explain it and its relationship to more traditional ways of quantifying potential-like energy functions of dynamical systems which we think the reviewer has in mind. We have now included a second type of energy landscape analysis which gives a complementary perspective on the RNN dynamics and is more straightforwardly comparable to typical potential functions. We describe the new analysis in the section "Large-scale neural predictions of recurrent neural network model" as follows:

"Crucially, there are two complementary viewpoints from which we can construct an energy landscape; the first allocentric (i.e., third-person view) perspective quantifies the energy associated with each position in state space, whereas the second egocentric (i.e., first person view) perspective quantifies the energy associated relative changes independent of the direction of movement or the location in state space. The allocentric perspective is straightforwardly comparable to the potential function of a dynamical system but can only be applied to low dimensional data in settings where a position-like quantity is meaningfully defined. The egocentric perspective is analogous to taking the point of view of a single particle in a physical setting and quantifying the energy associated with movement relative to the particles initial location. An egocentric framework is thus more applicable, when signal magnitude is relative rather than absolute. See materials and methods, and (see Fig S4 for an intuitive explanation of the allocentric and egocentric energy landscape analysis on a toy dynamical system)."

From the allocentric perspective it is entirely true that increasing gain increases the depth of the landscape, equivalent to increasing the depth of the attractor. However, because the input to the network changes dynamically the location of the approximate fixed-point attractor changes and the network state "chases" this attractor over the course of the trial. Importantly, the location of the energy minima changes more rapidly as gain increases, effectively forcing the network to rapidly change course at the point of the perceptual switch (see Fig 4). To quantify this effect we constructed a new measure - neural work - which describes the amount of "force" exerted on the low-dimensional neural trajectory by the vector field quantified by the allocentric landscape. Specifically we treat the allocentric landscape as analogous to a potential function and then leverage the fact that force is equal to the negative gradient of potential energy to calculate the work (force x displacement) done on the low dimensional trajectory at each time point. This showed that as gain increases the amount of work done on the neuronal trajectory at turning points increases analogous to the application of an external force transiently increasing the kinetic energy of an object. From the perspective of the egocentric landscape this results in a flattening of the landscape as there is

a lower energy (i.e. higher probability) assigned to large deviations in the neuronal trajectory around the perceptual switch.

Because of the novelty of the analyses we went to great lengths to carefully explain the methods in the updated manuscript. In addition we wrote a short tutorial style MATLAB script implementing both the allocentric and egocentric landscape analysis on a toy dynamical system with a known potential function (a supercritical pitchfork bifurcation).

> *(4) From the RNN model it is not clear how changes in excitatory and inhibitory gain lead to slower/faster switching. In order to better understand the role of inhibitory and excitatory gain on switching, I would suggest studying a simple discrete attractor model (a rate model, for example as in Wong and Wang 2006 or Roxin and Ledberg, Plos Comp. Bio 2008) which will allow to study these effects in terms of a very few model parameters. The Roxin paper also shows how to map rate models onto simplified one-dimensional systems such as the one in Fig S3. Setting up the model using this framework would allow for making much stronger, principled statements about how gain changes affect the energy landscape, and under which conditions increased inhibitory gain leads to faster switching.*
>
> *One possibility is that increasing the excitatory gain in the RNN leads to saturated firing rates. If this is the reason for the different effects of excitatory and inhibitory gain changes, it should be properly explained. Moreover, the biological relevance of this effect should be discussed (assuming that saturation is indeed the explanation).*

We thank the reviewer for this excellent suggestion. After some consideration we decided that studying a reduced model would likely not do justice to the dynamical mechanisms of RNN especially after making gain dynamical rather than stationary. Still we very much share the reviewer's concern that we need a stronger link between the (now dynamical) gain alterations and energy landscape dynamics. To this end we now describe and interrogate the dynamics of the RNN at a circuit level through selectivity and lesion based analyses, at a population level through analysis of the dynamical regime traversed by the network, and finally, through an extended energy landscape framework which has far stronger links to traditional potential based descriptions of low-dimensional dynamical systems (also see to comment 3. above).

At a circuit level the speeding of perceptual switches is mediated by inhibition of the initially dominant population we describe in paragraphs 7 and 8 of the section "Computational evidence for neuromodulatory-mediated perceptual switches in a recurrent neural network" as follows:

"Having confirmed our hypothesis that increasing gain as a function of the network uncertainty increased the speed of perceptual switches, we next sought to understand the mechanisms governing this effect starting with the circuit level and working our way up to the population level (c.f. Sheringtonian and Hopfieldian modes of analysis(66)). Because of the constraint that the input and output weights are strictly positive, we could use their (normalised) value as a measure of stimulus selectivity. Inspection of the firing rates sorted by input weights revealed that the networks had learned to complete the task by segregating both excitatory and inhibitory units into two stimulus-selective clusters (Fig 2C). As the inhibitory units could not contribute to the networks read out, we hypothesised that they likely played an indirect role in perceptual switching by inhibiting the population of excitatory neurons selective for the currently dominant stimulus allowing the competing population to take over and a perceptual switch to occur.

To test this hypothesis, we sorted the inhibitory units by the selectivity of the excitatory units they inhibit (i.e. by the normalised value of the readout weights). Inspecting the histogram of this selectivity metric revealed a bimodal distribution with peaks at each extreme strongly

inhibiting a stimulus selective excitatory population at the exclusion of the other (Fig S2). Based on the fact that leading up to the perceptual switch point both the input and firing rate of the dominant population are higher than the competing population, we hypothesized that gain likely speeds perceptual switches by actively inhibiting the currently dominant population rather than exciting/disinhibiting the competing population. We predicted, therefore, that lesioning the inhibitory units selective for the stimulus that is initially dominant would dramatically slow perceptual switches, whilst lesioning the inhibitory units selective for the stimulus the input is morphing into would have a comparatively minor slowing effect on switch times since the population is not receiving sufficient input to take over until approximately half way through the trial irrespective of the inhibition it receives. As selectivity is not entirely one-to-one, we expect both lesions to slow perceptual switches but differ in magnitude. In line with our prediction, lesioning the inhibitory units strongly selective for the initially dominant population greatly slowed perceptual switches (Fig 3F upper), whereas lesioning the population selective for the stimulus the input morphs into removed the speeding effect of gain but had a comparatively small slowing effect on perceptual switches (Fig 3F lower)."

At the population level we characterised the dynamics of the 2D parameter space (defined by gain and the difference between the input dimensions) traversed by the network over the course of a trial as input and gain dynamically change. We describe this paragraphs 9-14 of the section "Computational evidence for neuromodulatory-mediated perceptual switches in a recurrent neural network" which we reprint below for the reviewers convenience :

"Based on the selectivity of the network firing rates we hypothesised that the dynamics were shaped by a fixed-point attractor whose location and existence were determined by gain and and thus changed dynamically over the course of a single trial(67-70). Because of the large size of the network, we could not solve for the fixed points or study their stability analytically. Instead we opted for a numerical approach and characterised the dynamical regime (i.e. the location and existence of approximate fixed-point attractors) across all combinations of gain and visited by the network. Specifically, for each combination of elements in the parameter space we ran 100 simulations with initial conditions (firing rates) drawn from a uniform distribution between [0,1], and let the dynamics run for 10 seconds of simulation time (10 times the length of the task - longer simulation times did not qualitatively change the results) without noise. As we were interested in the existence of fixed-point attractors rather than their precise location, at each time point we computed the difference in firing rate between successive time points across the network. For each simulation we computed both the proportion of trials that converged to a value below 10^-2 giving us proxy for the presence of fixed points, and the time to convergence, giving us a measure of the "strength" of the attractor.

Across gain values when input had unambiguous values, the network rapidly converged across all initialisations (Fig 3A & 3C-H). When input became ambiguous, however, the dynamics acquired a decaying oscillation and did not converge within the time frame of the simulation. As gain increased, the range of values characterised by oscillatory dynamics broadened. Crucially, for sufficiently high values of gain, ambiguous values transitioned the network into a regime characterised by high amplitude inhibition-driven oscillations (Fig 3D & 3G). Each trial can, therefore, be characterised by a trajectory through this 2-dimensional parameter space, with dynamics shaped by the dynamical regimes of each location visited (Fig 3A-B).

When uncertainty has a small impact on gain the network has a trajectory through an initial regime characterised by the rapid convergence to a fixed point where the population representing the initial stimulus dominated whilst the other was silent (Fig 3C), an uncertain regime characterised by oscillations with all neurons partially activated (Fig 3D), and after passing through the oscillatory regime, the network once again enters a new fixed-point

regime where the population representing the initial stimulus is now silent and the other is dominant (Fig 3E).

For high gain trails, the network again started and finished in states characterised by a rapid convergence to a fixed point representing the dominant input dimension (Fig 3F-H), but differed in how it transitioned between these states. Uncertain inputs now generated high amplitude oscillations with the network flip-flopping between active and silent states (Fig 3G). We hypothesised that, within the task, this has the effect of silencing the initially dominant population, and boosting the competing population. To test this we initialised each network with parameter values well inside the oscillatory regime (u = [.5, .5] , gain = 1.5) with initial conditions determined by the selectivity of each unit. Excitatory units selective for input dimension 1, as well as the associated inhibitory units projecting to this population, were fully activated, whilst the excitatory units selective for input dimension 2 and the associated inhibitory units were silenced. As we predicted, when initialised in this state the network dynamics displayed an out of phase oscillation where the initially dominant population was rapidly silenced and the competing population was boosted after a brief delay (219 (ms), +/-114 Fig S3)."

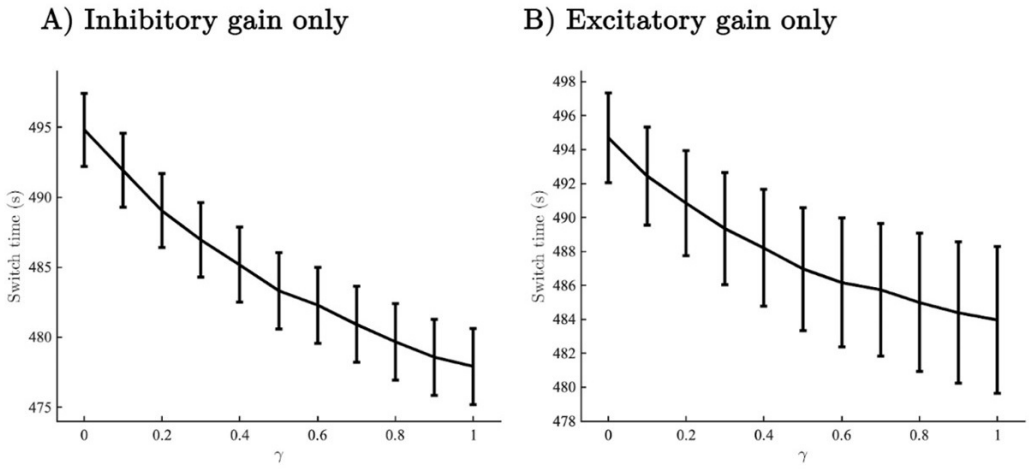
From this we concluded that at a population level, heightened gain leading up to the perceptual switch speeds the switch by transiently pushing the dynamics into an unstable dynamical regime replacing the fixed-point attractor representing the input with an oscillatory regime that actively inhibits the currently dominant population and boosts the competing population before transitioning back into a regime with a stable (approximate) fixed-point attractor representing the new stimulus (Fig 3F-H & Fig S3).

As we describe in the our response to comment 3 above our extended energy-landscape analysis framework now includes an explicit link between the potential of the dynamical system and allocentric landscape, whilst also explaining how a transient deepening of the allocentric landscape (which can be essentially thought of analogous to a traditional potential function) relates to the flattening of the egocentric landscape.

Finally, whilst we appreciate the interest in further characterising the effect of inhibitory gain compared with excitatory gain the topic is is largely orthogonal the aims of our paper so we have removed the discussion of inhibitory vs excitatory gain. Still, we understand that we need to do our due diligence and check that our results do not break down when we manipulate either inhibitory or excitatory gain in isolation. To this end we checked that dynamical gain still speeded perceptual switches when the effect was isolated to inhibitory or excitatory cells in isolation. We show the behavioural plots below for the reviewer's interest.

Author response image 1.

Switch time as a function of uncertainty forcing



Alternative mechanisms:

It is mentioned in the introduction that changes in attention could drive perceptual switches. A priori, attention signals originating in the frontal cortex may be plausible mechanisms for perceptual switches, as an alternative to LC-controlled gain modulation. Does the observed fMRI dynamics allow us to distinguish these two hypotheses? In any case, I would suggest including alternative scenarios that may be compatible with the observed findings in the discussion.

We agree with the reviewer, in that attention is itself a confound and a process that is challenging to disentangle from the perceptual switching process in the current task. Importantly, we were not arguing for exclusivity in our manuscript, but merely testing the veracity of the hypothesis that the ascending arousal system may play a causal role in mediating and/or speeding perceptual switches. Future work with experiments that more specifically aim to dissociate these different features will be required to tease apart these different possibilities.

> ***Reviewer #2 (Public Review):***
>
> *Strengths*
>
> *- the study combines different methods (pupillometry, RNNs, fMRI).*
>
> *- the study combines different viewpoints and fields of the scientific literature, including neuroscience, psychology, physics, dynamical systems.*
>
> *- This combination of methods and viewpoints is rarely done, it is thus very useful.*
>
> *- Overall well-written.*
>
> *Weaknesses*
>
> *- The study relies on a report paradigm: participants report when they identify a switch in the item category. The sequence corresponds to the drawing of an object being gradually morphed into another object. Perceptual switches are therefore behaviorally relevant, and it is not clear whether the effect reported correspond to the perceptual switch per se, or the detection of an event that should change behavior (participant press a button indicating the perceived category, and thus switch buttons when they identify a perceptual change). The text mentions that motor actions are controlled for, but this fact only indicates that a motor action is performed on each trial (not only on the switch*

> *trial); there is still a motor change confounded with the switch. As a result, it is not clear whether the effect reported in pupil size, brain dynamics, and brain states is related to a perceptual change, or a decision process (to report this change).*

We agree with the reviewer that the coupling of the motor change with the perceptual switch is confounded to some degree, but since motor preparation occurs on every trial we suspect that it is more accurate to describe it as confounded with task-relevance more than motor preparation per se. While it is possible that pupil diameter, network topology and energy landscape features are all related to motor change rather than the perceptual switch, we note that the weight of evidence is against this interpretation, given the simple mechanistic explanation created by the coupling of perceptual uncertainty to network gain.

> *- The study presents events that co-occur (perceptual switch, change in pupil size, energy landscape of brain dynamics) but we cannot identify the causes and consequences. Yet, the paper makes several claims about causality (e.g. in the abstract "neuromodulatory tone … causally mediates perceptual switches", in the results "the system flattening the energy landscape … facilitated an updating of the content of perception").*

We have made an effort to soften the causal language, where appropriate. In addition, we note that we have changed the title to “Gain neuromodulation mediates task-relevant perceptual switches: evidence from pupillometry, fMRI, and RNN Modelling” to reflect the fact that our claims do not extent to cases of perceptual switches where the stimulus is only passively observed.

> *- Some effects may reflect the expectation of a perceptual switch, rather than the perceptual switch per se. Given the structure of the task, participants know that there will be a perceptual switch occurring once during a sequence of morphed drawings. This change is expected to occur roughly in the middle of the sequence, making early switches more surprising, and later switches less surprising. Differences in pupil response to early, medium, and late switches could reflect this expectation. The authors interpret this effect very differently ("the speed of a perceptual switch should be dependent on LC activity").*

The task includes catch trials designed to reduce the expectation of a perceptual switch. In these trials, a perceptual switch occurs either earlier or later than usual. While these trials are valuable for mitigating predictability, we did not focus extensively on them, as they were thoroughly discussed in the original paper. Additionally, due to the limited number of catch trials, it is difficult—if not impossible—to calculate a reliable mean surprise per image set.

It is also worth noting that the pupil study does not include catch trials, which could contribute to differences in how perceptual switches are processed and interpreted between the fMRI and pupil experiments.

> *- The RNN is far more complex than needed for the task. It has two input units that indicate the level of evidence for the two categories being morphed, and it is trained to output the dominant category. A (non-recurrent) network with only these two units and an output unit whose activity is a sigmoid transform of the difference in the inputs can solve the task perfectly. The RNN activity is almost 1-dimensional probably for this reason. In addition, the difficult part of the computation done by the human brain in this task is already solved in the input that is provided to the network (the brain is not provided with the evidence level for each category, and in fact, it does not know in advance what the second category will be).*

We agree that a simpler model could perform the task. We opted to use an RNN rather than hand craft a simpler model as we wanted to use the model as both a method of hypothesis

testing and hypothesis generation. We now expand on and justify this modelling choice in the second paragraph of the discussion (also see our response to Reviewer 1 comment 4):

“We chose to use an RNN, instead of a simpler (more transparent) model as we wanted to use the RNN as a means of both hypothesis generation and hypothesis testing. Specifically, unlike more standard neuronal models which are handcrafted to reproduce a specific effect, when building an RNN the modeller only specifies the network inputs, labels, and the parameter constraints (e.g. Dale’s law) in advance. The dynamics of the RNN are entirely determined by optimisation. Post-training manipulations of the RNN are not built in, or in any way guaranteed to work, making them more analogous to experimental manipulations of an approximately task-optimal brain-like system. Confirmatory results are arguably, therefore, a first steps towards an in vitro experimental test.”

In other words, a simpler model would not have been appropriate to the aims. In addition we note that low dimensional dynamics are extremely common in the RNN literature and are in no way unique to our model.

> *- Basic fMRI results are missing and would be useful, before using elaborate analyses. For instance, what are the regions that are more active when a switch is detected?*

We explicitly chose to not run a standard voxelwise statistical parametric approach on these data, as the results were reported extensively in the original study (Stottinger et al., 2018).

> *- The use of methods from physics may obscure some simple facts and simpler explanations. For instance, does the flatter energy landscape in the higher gain condition reflect a smaller number of states visited in the state space of the RNN because the activity of each unit gets in the saturation range? If correct, then it may be a more straightforward way of explaining the results.*

We appreciate the reviewer's concern as this would indeed be a problem. However, this is not the case for our network. At the time point of the perceptual switch where the egocentric landscape dynamics are at their flattest the RNN firing rates are approximately 50% activated nowhere near the saturation point. In addition, a flatter landscape in the egocentric and allocentric landscape analyses only occurs - mathematically speaking - when there are more states visited not less.

In addition, we note that we are very sympathetic to the complexity of our physics based analyses and have gone to great lengths to describe them in an accessible manner in both the main text and methods. We have also included tutorial style code demonstrating how the analysis can be used on a toy dynamical system in the supplementary material.

> *- Some results are not as expected as the authors claim, at least in the current form of the paper. For instance, they show that, when trained to identify which of two inputs u1 and u2 is the largest (with u2=1-u1, starting with u1=1 and gradually decreasing u1), a higher gain results in the RNN reporting a switch in dominance before the true switch (e.g. when u1=0.6 and u2=0.4), and vice et versa with a lower gain. In other words, it seems to correspond to a change in criterion or bias in the RNN's decision. The authors should discuss more specifically how this result is related to previous studies and models on gain modulation. An alternative finding could have been that the network output is a more (or less) deterministic function of its inputs, but this aspect is not reported.*

We appreciate this comment but it is simply not applicable to our network. There is no criterion in the RNN. We could certainly add one but this would be a significant departure from how decisions are typically modelled in RNNs. The (deterministic) readout is the max of the projection of the (instantaneous) excitatory firing rate onto the readout weights. A shift in

criterion would imply that the dynamics are unaffected and the effect can be explained by a shift in the readout weights; this cannot be the case because the readout weights are stationary the change occurs at the level of the activation function.

We are aware that there is a large literature in decision making and psychophysics that uses the term gain in a slightly different way. Here we are strictly referring to the gain of the activation function. Although we agree that it would be interesting and important to discuss the differing uses of the term gain, this is beyond the scope of the present paper.

https://doi.org/10.7554/eLife.93191.2.sa0